BEHAVIORAL AND BRAIN SCIENCES (1990) 13, 109–161
Printed in the United States of America

Insensitivity of the analysis of variance to heredity-environment interaction

Douglas Wahlsten*
Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G2E9
Electronic mail: *userdlwa@ualtamts.bitnet*

Abstract: It makes sense to attribute a definite percentage of variation in some measure of behavior to variation in heredity only if the effects of heredity and environment are truly additive. Additivity is often tested by examining the interaction effect in a two-way analysis of variance (ANOVA) or its equivalent multiple regression model. If this effect is not statistically significant at the $\alpha = 0.05$ level, it is common practice in certain fields (e.g., human behavior genetics) to conclude that the two factors really are additive and then to use linear models, which assume additivity. Comparing several simple models of nonadditive, interactive relationships between heredity and environment, however, reveals that ANOVA often fails to detect nonadditivity because it has much less power in tests of interaction than in tests of main effects. Likewise, the sample sizes needed to detect real interactions are substantially greater than those needed to detect main effects. Data transformations that reduce interaction effects also change drastically the properties of the causal model and may conceal theoretically interesting and practically useful relationships. If the goal of partitioning variance among mutually exclusive causes and calculating "heritability" coefficients is abandoned, interactive relationships can be examined more seriously and can enhance our understanding of the ways living things develop.

Keywords: causal models; gene action; heritability; nature/nurture; power; sample size; scale transformation

1. Introduction

The statistical analysis of data helps the researcher detect consistent patterns of results that might otherwise be obscured by uncontrolled and unknown sources of variation. Like every analytical technique, a statistical method is based on certain assumptions about the properties of the objects being studied. If assumptions are not valid, the method can lead to erroneous conclusions just as readily as can a faulty laboratory procedure. A method can be used with confidence only if there are effective ways to test its validity. As discussed by Crusio (in press), certain experimental designs do not lend themselves to tests of crucial assumptions, no matter how many observations are made. Another difficulty, the focus of this target article, arises when a test is possible, at least in principle, but is so insensitive that violations of assumptions often escape detection.

The widespread application of the analysis of variance (ANOVA) to factorial experiments in the behavioral and brain sciences provides a case in point. This technique, first devised by Fisher and Mackenzie (1923) for use in agriculture, is convenient for evaluating the results of an experiment in which every category of one factor (e.g., variety of a crop species) is combined with every condition of another factor (e.g., kind or amount of fertilizer). The classical ANOVA method is gradually being replaced by a more flexible technique, multiple regression, which fits data to a linear equation with one term for each separate "effect" in the experiment, but for simple factorial designs the two procedures are essentially the same (Edwards 1979). ANOVA partitions the total variation in a measure (e.g., crop yield) among four contributing causes: (a) the "main" effect of variety averaged over all kinds of fertilizer, (b) the main effect of fertilizer averaged over all varieties, (c) the "interaction" of variety and fertilizer, and (d) sources of variation or "error" within each group. Interaction in a factorial experiment signifies the departure of a group mean score from the simple sum of the respective main effects. If present, it indicates that crop yield depends on the specific *combination* of variety and treatment. One of the great merits of the ANOVA method is that it can readily detect interaction. Unfortunately, the technique is relatively insensitive to certain types of interaction and can be quite misleading when interpreted uncritically.

Many psychological theories rise or fall with the occurrence or absence of statistical interaction. Discussing the question of whether or not drive and reinforcement are independent, Mackintosh (1974) wrote: "In principle, the question should be answered easily, requiring no more than a large factorial experiment in which several levels of drive are combined with several magnitudes of reinforcement, with an analysis of variance being performed to test for a significant interaction of the two factors" (p. 154). Another example is the "person-situation" question. Psychologists ask whether an individual has a distinct personality, which remains the same in a

 0140-525X/90 $5.00+.00

variety of situations (relative stability), or whether personality is highly flexible and specific to circumstance (situationism). It could also happen that personality changes according to the situation but that the kind of change depends on stable characteristics of the person (coherence). Rival explanations such as relative stability and coherence are often contrasted using ANOVA. Magnusson and Allen (1983) state: "Though most of the variance in a Person × Situation matrix of data is usually because of the main effects of persons, enough variance is left that can be explained by interindividual differences in patterns of cross-situational profiles to support the coherence model" (p. 24).

The detection and interpretation of interaction are important in virtually every area of the behavioral and brain sciences, but they are nowhere so crucial as in human behavior genetics, where the prevailing models seek to partition variance between two sources, nature and nurture. Controversies in behavior genetics (e.g., Henderson 1979; Wahlsten 1979) have led to further questions about the validity and sensitivity of analysis of variance. The answers have implications for many other fields of study. The following discussion is therefore directed to a specific issue in behavior genetics but can easily be extended far beyond behavior genetics.

2. Two research agendas

Almost any characteristic of living organisms can be shown to vary as a consequence of both heredity and environment. Some studies attempt to understand the causes of these individual differences by examining the functional roles of heredity and environment in individual development, especially how they relate to or depend upon each other; other studies try to estimate the strength of the influence of one factor versus the other in a population of organisms. These two research agendas can be contradictory. It is possible to ascribe a definite percentage of individual differences in a population to variation in heredity, for example, only if heredity and environment are strictly additive and act separately from one another in the course of development.

Let us recall a theorem from introductory statistics. If one variable, Y, is the sum of two other variables, X and Z, then the variance of Y is equal to:

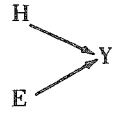
$$\mathrm{Var}(Y) = \mathrm{Var}(X) + \mathrm{Var}(Z) + 2\mathrm{Cov}(XZ).$$

If X and Z are uncorrelated, then Cov(XZ) = 0 and the variance of a sum is the sum of the separate variances. Suppose X is a measure of one's heredity (H) and Z represents one's environment (E). We then arrive at the basic causal model in quantitative behavior genetics, Y = H + E, according to which a measured characteristic of an individual is the sum of the two separate components. This model is the conceptual basis for analysing or partitioning variance in a population, because it implies that:

$$\mathrm{Var}(Y) = \mathrm{Var}(H) + \mathrm{Var}(E).$$

As expressed by Fuller and Thompson (1978) (who used P for "phenotype" rather than Y): "The fundamental problem of quantitative behavior genetics is to partition V_p [Var(Y) here}] into its components so as to estimate the proportional contributions of genes and life histories to population variability" (p. 52). The heritability coefficient (h^2) in the broader sense expresses the proportion of measured variation among individuals attributable to variation in their heredities, Var(H)/Var(Y). According to Plomin (1988): "Behavioral genetics is only useful for addressing the extent to which genetic and environmental variation contribute to phenotypic variation in a population" (p. 107).

The Y = H + E model corresponds to a very simple diagram of causal relations:



This implies that the influence of heredity on the development and eventual magnitude of some characteristic is completely separate and distinct from the influence of environment, and that the effect of environment does not depend on a person's heredity. Heritability analysis, to be valid, requires that a particular model of development be true. Research results that cast doubt on the additivity of H and E necessarily cast doubt on the interpretation of heritability (McGuire & Hirsch 1977; Wahlsten 1979) because nonadditivity of the contributing causes makes it invalid to partition the variance into distinct components and thereby renders a heritability coefficient meaningless (Lewontin 1974). Quantitative genetics broadly conceived can incorporate interactive effects (e.g., Cavalli-Sforza & Feldman 1973), but heritability analysis cannot.

The direct investigation of individual development through longitudinal observation and concurrent experimental manipulation of heredity and environment, on the one hand, makes no a priori assumption about the additivity of H and E. Rather, it uses genetic variants to help interrogate nature. Logically, this research agenda ought to precede attempts to partition variance but, historically, it did not. Academic interest in hereditary sources of individual differences in intelligence and other human attributes preceded the scientific study of behavioral development by many years (Fancher 1985), just as statistical techniques designed to partition variance predated important insights into the roles of genes in development.

3. Developmental and statistical interactions

Because heritability analysis requires the absence of statistical interaction involving H and E, and because the basic formula Y = H + E is a model for an individual, the question of statistical interaction is sometimes posed as a question of whether H and E interact in the course of development. This can lead to some confusion in terminology and meaning because differing interpretations of interaction and "interactionism" abound, especially among psychologists.

In personality theory, for example, "interactionism" is sometimes taken to mean that the combined effects of the qualities of individuals and the situation in which they are reared or tested must be considered (Bowers 1973; Magnusson & Allen 1983), which to some theorists makes the interaction term in ANOVA of critical importance. However, if one simply asserts that *both* factors must be considered, this does not necessarily invalidate an ad-

ditive model (H + E). On the contrary, it can lead to bold claims that quantitative genetic analysis will finally resolve the person-situation debate (Rowe 1987).

Concerning human intelligence, Fancher (1985) writes: "Everyone now recognizes that heredity and environment never work in isolation, but only in *interaction* with each other. From the moment of birth onwards, each child's real or presumed 'nature' helps determine its nurture" (p. 231), as when a "bright" child is given special advantages. He hopes scientists will achieve "an approximate appreciation of the relative strengths of the two factors." Evidently, Fancher uses "interaction" in the sense of the *covariance* of H and E, which remains compatible with additivity.

Another concept of interaction is the genetically determined "norm of reaction" (Platt & Sanislow 1988), in which the kind and degree of response of a developing organism to a particular environment is itself assumed to be hereditary. This notion, advocated strongly by Schmalhausen (1949), is generally not compatible with the additivity of effects of H and E in a factorial experiment (Lewontin 1974). However, Schmalhausen definitely separated the causal contributions of H and E: "In the development of any individual, environmental factors act only as agents releasing form building processes and providing conditions necessary for their realization" (p. 2).

As Oyama (1985) has so well documented, many contemporary advocates of interactionism assign a one-sided role to the genotype as the source of information giving *form* to living things. Her own use of interactionism is fundamentally different. [See also Johnston: "Developmental explanation and the ontogeny of birdsong" *BBS* 11 (4) 1988.] The form of a developing organism is seen as a product of the interactions among the parts of the system, so that "the informational function of any developmental interactant is dependent on the rest of the system" (Oyama 1988, p. 99). If there is no developmental information inherent in a component of a living thing apart from its multifarious relations with other componenets, it makes no sense to assign a certain fraction of a phenomenon to one contributing cause. However, for Oyama (1988, p. 98): "Interactionism does not dictate any particular outcome" of a study, and it does not require that statistical interaction be observed in every experiment.

Thus, some views of developmental interaction (e.g., Fancher's) are compatible with the additivity of H and E, whereas others (Schmalhausen's) assume nonadditivity, and yet another (Oyama's) makes no consistent prediction of statistical results. Finding a statistical interaction between H and E would place heritability analysis in peril but would not by itself allow us to draw finer distinctions between the norm of reaction and Oyama's interactionism.

4. Testing alternative models

A perusal of the current literature indicates that the heritability coefficient and the general idea of partitioning variance are very widely accepted in behavioral science. A rather small number of scholars may be aware of the insecure foundations of this approach, but a large majority of interested readers is not. One objective of this target article, therefore, is to explain clearly why and how the presence of statistical interaction should be assessed. To understand this problem better, let us contrast two alternative models. The scientific method requires that, to demonstrate that one hypothesis is true, reasonable alternative hypotheses must be shown to be false.

Model I: $Y = H + E$
Model II: $Y = H \cdot E$

According to the second model, heredity and environment are multiplicative rather than additive. This means that an individual with a heredity more favorable for or vulnerable to developing some characteristic will change more in response to a particular change in environment than will one with a lower value of H. For example, the induction of neural tube defects by various doses of insulin given to pregnant mice occurs with a steeper dose-response curve in fetuses carrying the genes rib fusion (Rf) or crooked (Cd) than in their littermates (Cole & Trasler 1980). Many other reasonable models of H × E interaction could be postulated (e.g., Cavalli-Sforza & Feldman 1973), but this one is the simplest and reveals the fundamental difficulty with heritability analysis. A multiplicative model also provides good expression of a deeper meaning of interaction, as with the formula for the area of a triangle, where it makes no sense to assign greater responsibility for the area to the length of the base than to the height or vice-versa.

Let us use these models to predict the outcomes of some simple experiments we could do in a laboratory. Let H_j represent the effect of the heredity of a particular strain of animal and let E_k represent the effect of the environment in which it is raised. Of course, genes themselves are sequences of nucleotide bases in DNA molecules and as such are categorical variables, whereas H_j is taken to be a continuous variable on an interval scale. For purposes of explication here, the usual procedure of quantitative genetics (Plomin et al. 1980) is used, which maps genotypes at many loci onto a single scale of measure. The value Y_{ijk} is a measure of an individual i with heredity j reared in environment k, and M_{jk} is the mean score of a large number of such organisms.

For our first experiment, let us raise equal numbers of mice of strain 1 in two different environments, which is the proper method for assessing the plasticity or modifiability of a characteristic. The design has only two groups.

	E_1	E_2
H_1	M_{11}	M_{12}

It seems intuitively obvious that any difference in group means, $\Delta M = M_{11} - M_{12}$, must be attributable solely to the difference in environment, because all subjects have the same heredity. This may be reasonable logically, but it is mathematically true in general only if Model I is correct (or if the functions for the two strains are parallel across the range of E_k). Now, compare predictions of the two models.

Model I: $\Delta M = (H_1 + E_1) - (H_1 + E_2)$
$= (H_1 - H_1) + (E_1 - E_2) = \Delta E$

According to Model I, the group mean difference does not depend on which strain we choose for the experiment.

Model II: $\Delta M = H_1 E_1 - H_1 E_2 = H_1(E_1 - E_2) = H_1 \Delta E$

According to Model II, the group difference depends jointly on the magnitude of the environmental difference (ΔE) *and* the strain's heredity. The larger the magnitude of H_1, the greater will be the group difference, which is a clear case of interaction.

Next, let us compare two strains raised in the same laboratory environment, which allows a crude measure of heritability (Hegmann & Possidente 1981).

	H_1	H_2
E_1	M_{11}	M_{21}

Compare the predictions of the two models:

Model I: $\Delta M = (H_1 + E_1) - (H_2 + E_1)$
$= (H_1 - H_2) + (E_1 - E_1) = \Delta H$

Model II: $\Delta M = H_1E_1 - H_2E_1 = E_1(H_1 - H_2) = E_1\Delta H$

Again intuition tells us that ΔM must reflect only ΔH, but according to the multiplicative model the manifestation of a particular strain difference in heredity depends on the environment in which the animals are raised. Under Model II, the proportion of total variance attributable to the strain difference is no longer a valid indication of the magnitude of the ΔH effect or of "heritability" in the usual sense.

Although the two models make very different numerical predictions for both experiments, there is no practical way to test them because there is no way to measure the H or E component directly. Lacking this, both models predict that the group difference will not be zero, and virtually any outcome is consistent with either model. Hence, an experiment must be designed so that the two models predict distinctly different testable outcomes. The solution is a two-way factorial experiment, in which at least two strains are reared in at least two environments.

	H_1	H_2
E_1	M_{11}	M_{21}
E_2	M_{12}	M_{22}

We can ask whether the difference in strain means in E_1 is the same as in E_2.

Model I: ΔM in $E_1 = (H_1 + E_1) - (H_2 + E_1) = H_1 - H_2 = \Delta H$
ΔM in $E_2 = (H_1 + E_2) - (H_2 + E_2) = H_1 - H_2 = \Delta H$

Therefore,

$$(\Delta M \text{ in } E_1) - (\Delta M \text{ in } E_2) = \Delta H - \Delta H = \boxed{0.}$$

Model II: ΔM in $E_1 = H_1E_1 - H_2E_1 = E_1\Delta H$
ΔM in $E_2 = H_1E_2 - H_2E_2 = E_2\Delta H$

Therefore,

$$(\Delta M \text{ in } E_1) - (\Delta M \text{ in } E_2) = E_1\Delta H - E_2\Delta H = \boxed{\Delta H \Delta E}$$

Model I predicts that the strain difference will be the same in both environments, whereas Model II predicts they will be different. The usual way to evaluate these alternatives is two-way analysis of variance (ANOVA), in particular the interaction term. The additive model requires that there be no significant interaction between strain and environment, whereas Model II and a host of other models expect significant interaction. This is essentially the same test proposed by Plomin et al. (1977) for use in human adoption studies. They noted that the test proposed by Jinks and Fulker (1970) using monozygotic twins "may confound some purely environmental effects with genotype-environment interaction" (p. 314). Furthermore, Vetta (1981) has pointed out a serious algebraic error in the Jinks and Fulker (1970) paper which renders their test of interaction meaningless.

If there is agreement about this general approach for assessing H × E interaction, what are the results of its use in practice? Even among specialists in behavioral genetics there is still widespread support for Plomin's (1988) view that H and E are additive and that behavioral genetics "is only useful" for partitioning variance. Here the problem is not a lack of understanding about the importance of interaction in theory. Rather, there is a divergence of opinion about its occurrence in reality. A central issue in this regard is the sensitivity of the test of additivity to the presence of real nonadditivity in the data.

Interaction has been evaluated in studies of human IQ and usually none is seen (Plomin et al. 1977; Plomin & DeFries 1983; Plomin 1986). Generalizations have then been made that heredity and environment are truly additive, and sophisticated path models have been derived to partition variance and covariance under the assumption that interaction is negligible (e.g., Heath et al. 1985; Henderson 1982; Phillips et al. 1987; Plomin et al. 1985). On the other hand, an immense collection of well-controlled laboratory studies of animals has provided abundant evidence of significant and illuminating interactions between heredity and environment (Carlier & Nosten 1987; Cole & Trasler 1980; Erlenmeyer-Kimling 1972; Goodall & Guastavino 1986; Kinsley & Svare 1987). At the 1987 Behavior Genetics Association meeting in Minneapolis, the concurrent sessions on human and animal studies were almost like two separate worlds in terms of attitudes towards interaction. Many human behavior geneticists dismissed interaction and cited heritability estimates with great confidence, while most of those studying mice, rats, and fruit flies documented one case of interaction after another and expressed skepticism about heritability coefficients.

How can it be that investigators draw such different conclusions about heredity-environment interaction? It is argued in this target article that the commonplace tests of interaction using ANOVA (analysis of variance) are relatively insensitive or have relatively low power to detect nonadditivity. The usual practice is to hypothesize zero interaction and, if no significant interaction term is found, to conclude that the factors are truly additive, which is tantamount to accepting a null hypothesis of additivity as true. In research with laboratory animals where heredity is under experimental control and large numbers of subjects with the same genotype can be

assigned to rearing in distinctly different environments, substantial interactions are often detected, whereas they may pass unseen in an adoption study because of low power of the statistical test. The history of this problem suggests that serious errors of interpretation can occur if ANOVA is applied uncritically.

5. The problem of power: History

What has been termed an "unpleasantness" about the analysis of variance of a factorial design (Traxler 1976) was first pointed out by Neyman (1935) in response to a presentation on the topic by Yates (1935) at a meeting of the Royal Statistical Society in England. Yates touted the factorial design as a method for detecting interactions, yet he stated that "if there is no evidence of interaction . . . the two factors . . . may be regarded as additive" (p. 193). Neyman responded with a hypothetical numerical example in which applying certain fertilizers a, b, and c to a plot separately reduced yields but in several combinations increased yields. He then used Monte Carlo simulation to generate 30 random sets of data from his hypothetical population, obtaining 27 main effects of a significant at the 0.01 level but nine instances when the a main effect was significant while neither the a × b nor a × c interaction achieved significance at even the 0.05 level. He warned that when interactions "do exist and are somewhat malicious the method may give unsatisfactory results" (p. 238), and he concluded that "the cause of the trouble lies in interactions which are very large and yet, owing to insufficient replication, are not likely to be found significant" (p. 241).

Tang (1938) used the noncentral *F* distribution to determine precisely the power of the one-way ANOVA. Kempthorne's (1952) influential treatise explained how to calculate power for the one-way design and suggested how to do it for interaction terms involving one degree of freedom. For other situations, he assured the reader: "It is a simple matter to obtain the sensitivity of any experiment" (p. 225). Kempthorne noted that the results of a factorial design may be "difficult to interpret" when interactions are appreciable with respect to main effects and warned the tests of additivity "may have rather low power in detecting non-additivity" (p. 258). Scheffé (1959) also gave examples of power for the one-way design and claimed that "calculations for other experimental designs are similar" (p. 62). He further advised that if additivity of factors is to be accepted on the basis of a nonsignificant test of interaction, "it is wise to try to answer the question whether this test has reasonable power" (p. 94). Thus, by 1960 the importance of the problem of interaction for ANOVA and the proper approach to calculating power were generally understood by expert statisticians, although the degree of insensitivity to interaction was not widely known.

The work of Cohen (1977) made power calculations for interactions more readily accessible to the less mathematically sophisticated in the behavioral and brain sciences. However, little use was made of this feature of the book. Reviewing the situation in 1976, Traxler observed that many modern experimenters interested in synergistic or other interactive effects seem to lack awareness of the problem of low power. Kraemer and Thiemann (1987) remarked recently that "those who are able to do power calculations readily are generally those who least know the fields of application, and those who best know the fields of application are least able to do power calculations" (p. 99). Their own work will help to overcome this problem, except with regard to interactions in ANOVA, which they did not discuss.

Today the problem of the power to detect interaction, which is certainly relevant for any research involving factorial design and ANOVA, is not generally understood among the practitioners or the consumers of behavior genetic research. From time to time there has been mention of the rather low power of tests of heredity-environment interaction (Eaves et al. 1977; Freeman 1973), but this has remained obscure in the pages of specialty journals. The present target article tries to explain this matter in a way that will make it comprehensible for anyone familiar with basic algebra and the ANOVA method.

6. An instructive example: Gravitation

A great danger in using ANOVA may occur when the true state of nature is markedly nonadditive but the statistical test is oblivious to this and misrepresents reality as additive. What if we apply factorial design and ANOVA to a situation known to be governed by a physical law? For example, according to Newton's law of universal gravitation, the force (F) of attraction between two objects is proportional to the product of their masses (m_1 and m_2) divided by the square of the distance (d) between their centers of mass. The G value in the equation is the gravitational constant.

$$F = \frac{Gm_1m_2}{d^2}$$

What would happen if a zealous advocate of heritability analysis were transported to a physics laboratory and asked to determine the nature of gravitation empirically? He might construct a simple apparatus as in Figure 1, where a 100 kg iron ball is affixed to a bench and another iron ball is suspended by a fine wire at a distance (d′) from the surface of the first ball. The displacement of the second ball by the first yields a measure of force. If our experimenter runs a study with four levels of mass (m_2) combined factorially with four distances (d′) between the balls, as in Table 1, the results for the ANOVA will be as shown in Table 2. The range of mass is limited by his ability to move the weights and the distance is limited by the size of the room. One presumes he makes small measurement errors on four trials under each condition of the study, resulting in a small within-group variance, so that the actual means of four separate measures in each condition deviate somewhat from the theoretical values in Table 1.

The experimenter's conclusions from the ANOVA would be that both mass and distance are important for the force, although the internal factor (mass) is rather more important and accounts for more variance than the external factor (distance), and that mass and distance are additive because the interaction term is not even close to significance. He might even proclaim a simplified law of

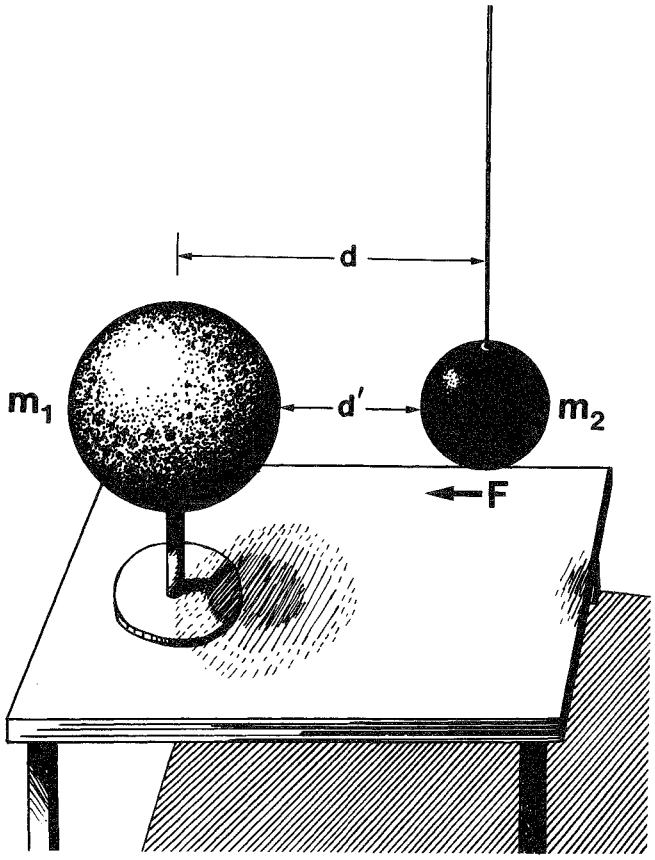


Figure 1. Apparatus to measure the force of gravitational attraction (F) between a large iron ball (m_1) and a suspended iron ball (m_2) whose surfaces and centers of gravity are d′ and d cm apart, respectively.

gravitation, F = μ + m + d. No doubt the interaction would have been significant, had a wider range of mass and distance been observed with more replications of the experiment. Indeed, a really large sample analyzed with the more sophisticated techniques of multiple regression and factor analysis can lead to models of motion far more complicated than anything ever imagined by Newton (I. Nabi, cited in Levins & Lewontin 1985).

7. Insensitivity to H × E interaction

Insensitivity to nonadditivity is not specific to the gravitation example. It is inherent in the typical use of the analysis of variance procedure, because ANOVA regards interaction as whatever is left over after the main effects of each factor averaged over all levels of the other factor(s)

Table 1. *Expected force of attraction (dynes)*

	Distance (d′)			
Mass (m_2)	100	125	150	175 cm
25	.0109	.0076	.0055	.0042
50	.0210	.0146	.0108	.0083
75	.0307	.0215	.0159	.0122
100 Kg	.0401	.0281	.0208	.0160

Table 2. *Two-way* ANOVA *for data in Table 1*

Source	SS	df	MS	F	est ω^2
Mass	0.00355	3	0.00118	12.97**	0.28
Distance	0.00246	3	0.00082	8.98**	0.19
Mass x Distance	0.00120	9	0.00013	1.46*	—
Error	0.00438	48	0.00009		

*$P > 0.10$
**$P < 0.0005$

have been taken into account (Fisher & Mackenzie 1923). To know just how insensitive it may be, one must calculate statistical power.

The power of a statistical test is the probability of rejecting a false null hypothesis. A particular hypothesis, such as additivity of heredity and environment, may fail to be rejected on the basis of ANOVA for one of two reasons: (a) It may be true. (b) It may be false, but the test may have low power. The degree of power of the test of one hypothesis can be assessed only with reference to a specific alternative hypothesis. Additivity must be judged with regard to specific kinds of nonadditivity.

If the additive model of behavior genetics (Y = H + E) predicts no significant H × E interaction, what is the power of a test of this hypothesis against the simple multiplicative model (Y = H·E)? This can be answered by supposing that the true relation is multiplicative and then determining what the results of an experiment would be. Suppose the score of an individual is $Y = H \cdot E + \epsilon$, where ε is the deviation of that individual from the mean of all those with the same heredity reared in the same environment. Let the values of ε be normally distributed with a mean of zero and variance σ^2. For simplicity, suppose the experiment is done with J strains reared in K different environments, and that the levels of H are equally spaced at h units apart and levels of E are e units apart. The score for individual i from strain j in environment k is taken to be

$$Y_{ijk} = (jh)(ke) + \epsilon_i,$$

and the expected value of all members of that group is

$$M_{jk} = (jk)(he).$$

From this relation we can easily determine the group means, as shown below.

ENVIRONMENT (k) \ STRAIN (j)	1	2	3	·····	J	M_k
1	he	2he	3he		Jhe	(J + 1)he/2
2	2he	4he	6he		2Jhe	2(J + 1)he/2
3	3he	6he	9he		3Jhe	3(J + 1)he/2
⋮						⋮
K	Khe	2Khe	3Khe		JKhe	K(J + 1)he/2

These expected means are all we need to calculate power of the tests of main effects and interaction. Cohen (1977) estimates power in terms of the effect size para-

meter (f) which is homologous to the effect size measure (d) for a *t* test on two independent groups:

$$d = \frac{M_1 - M_2}{\sigma}, \text{ for two groups;}$$

$$f = \frac{\sigma_M}{\sigma}, \text{ for J groups.}$$

The standard deviation σ is a measure of variation within a group, whereas σ_M is the standard deviation between true group means. Effect size compares differences between groups to variation within groups. The d coefficient denotes the number of standard deviations by which two true group means differ. Cohen (1977) considers d values of 0.2, 0.5 and 0.8 to represent small, medium, and large effect sizes, respectively, in psychological research. The f coefficient of effect size compares the standard deviation between several true group means to the standard deviation within a group.[1] Cohen (1977) considers f values of 0.1, 0.25, and 0.4 to be small, medium, and large effect sizes, respectively, in analysis of variance with several groups. Effect sizes tend to be smaller when there are several groups because some of the groups are likely to have intermediate values.

For K environments, the standard deviation of row means is defined as

$$\sigma_E = \sqrt{\frac{\sum_{k=1}^{K}(M_k - M)^2}{K}}$$

where M is the grand mean of all groups and M_k is the mean for environment k. In the case of a two-way factorial experiment where J strains are each reared in K different environments, the mean for environment k, M_k, is the average across the J strains. It follows that:[2]

$$\sigma_E = \frac{(J + 1)he}{4}\sqrt{\frac{(K + 1)(K - 1)}{3}}$$

Likewise for variation among the *J* strains,

$$\sigma_H = \frac{(K + 1)he}{4}\sqrt{\frac{(J + 1)(J - 1)}{3}}$$

When J = K, a square factorial design,

$$\sigma_H = \sigma_E = \frac{(J + 1)he}{4}\sqrt{\frac{(J + 1)(J - 1)}{3}}$$

The effect for interaction, σ_I, compares each group mean, M_{jk}, to the value expected from the sum of the main effects. That is, interaction in a two-way ANOVA is regarded as the "leftovers" after additive effects have been taken into account. For strain j reared in environment k, the mean value expected from the two separate main effects combined additively is

$$M + (M_j - M) + (M_k - M) = M_j + M_k - M,$$

and the deviation of the true group mean from this is

$$M_{jk} - (M_j + M_k - M) = M_{jk} - M_j - M_k + M.$$

Across all J·K groups,

$$\sigma_I = \sqrt{\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}(M_{jk} - M_j - M_k + M)^2}{JK}}$$

which yields (see Note 2)

$$\sigma_I = \frac{he}{12}\sqrt{(J + 1)(J - 1)(K + 1)(K - 1)}$$

When J = K:

$$\sigma_I = \frac{(J + 1)(J - 1)he}{12}$$

Now, for the purpose of calculating power, the principal concern is with power of tests of main effects relative to power of the test of interaction, which may be determined using the ratio f_H/f_I. For the multiplicative model with equal numbers of strains (*J*) and environments (*K*):

$$f_H/f_I = \frac{\sigma_H/\sigma}{\sigma_I/\sigma} = \sqrt{\frac{3(J + 1)}{(J - 1)}}$$

Thus, to compute power we can first specify f_H and then determine f_I from the above ratio. This is done in Table 3 for small, medium, large, and very large values of f when there are 10 subjects per group and $\alpha = 0.05$ (see Note 3).

Clearly, the test of H × E interaction when a multiplicative model obtains has very low power compared to the tests of main effects, which tells us that with n = 10 the ANOVA will usually point to additivity of H and E. As the number of strains and environments is made larger, the power of the test of interaction becomes greater, but even with 25 groups and 250 subjects it reaches only a modest 57%. If the Bonferroni correction is applied to the α level because several tests are being done simultaneously, the power of the tests of main effects and interaction will both decline but the problem of the relatively low power of the test of interaction will remain and could even be magnified for certain effect sizes and sample sizes.

The results for a 2 × 2 design may seem a little perplexing at first glance. After all, there will be one degree of freedom for the numerator and effective sample size of 19 (see Note 3) for the tests of main effects and interaction alike. Shouldn't the power functions for both main effects and interaction therefore be identical? Definitely not. The shape of the power function in ANOVA is indeed determined by the degrees of freedom, but it is also determined by the noncentrality parameter (Tang 1938), which is in turn determined by the effect size f (see Note 1). The principal problem of power in two-way ANOVA is not simply a matter of degrees of freedom.

Table 3. *Power of tests of main effects of H (and E) and H x E interaction using $\alpha = 0.05$ and n = 10 subjects per group, when Y = H·E. J = number of strains and environments.*

Main effect size	Test of strain effect				Test of interaction			
(f_H)	J = 2	3	4	5	J = 2	3	4	5
0.1	9	11	14	19	5	6	6	6
0.25	31	50	71	88	7	9	12	14
0.4	67	92	99	>99	12	18	26	36
0.5	87	99	>99	>99	16	27	41	57

Rather, it follows from the way the variance among all JK groups is partitioned among main effects and interaction, and this partition depends on the specific model of nonadditivity chosen as an alternative to the null hypothesis of additivity. There is no such thing as a power function existing apart from specific numerical alternatives to the null.

8. Reasonable alternatives to additivity

The simple multiplicative model is not the only reasonable alternative to additivity when different strains are involved. If the response is linear for each strain, there is no reason why the *Y* intercept should always be zero, as with Y = H·E. Consequently, several other models shown in Figure 2a were assessed for power of main effects and interaction. Certain of these were similar to models proposed for interacting genetic and cultural inheritance (Cavalli-Sforza & Feldman 1973) and for mental disorders (Kendler & Eaves 1986). Because the "norm of reaction" is often not linear when a wide range of environments is evaluated (Henry 1986), two nonlinear models were also considered. Although it is sometimes proposed that the norm of reaction is genetically determined (e.g., Hull 1945; Schmalhausen 1949; Via & Lande 1985), this is not realistic because the response to a new environment also depends on prior rearing conditions (Denenberg 1977). Nevertheless, for clarity, each model assumes that any parameters (a, b) are specified by heredity and that parameter values are equally spaced for the five strains. Rather than derive the ratio of effect sizes (f_H/f_I) using algebra, a computer program was written to generate expected means for a five strain by five environment (X = value of E) design and then to calculate σ_H, σ_E, and σ_I. Table 4 presents power estimates for each model when n = 10 and $\alpha = 0.05$. The largest main effect, be it for H or E, is taken to have a large effect size, f = 0.4.

In no case does the power of the test of interaction achieve an acceptable level of 80% or more when one or both main effects are virtually certain to be detected with ANOVA. It comes close to 80% for two Y = a + bX models, but when main effect size is 0.3 for these, the power of the test of main effects is 98% but the power of the test of H × E interaction is only 46%. The power of the test of interaction is relatively low even when the directions of effects of environment are opposite for several strains (Y = a + bX, Case 1, and $Y = aXe^{-bX}$), or when the rank orders of the strains change across environments (Y = a + bX, Case 2).

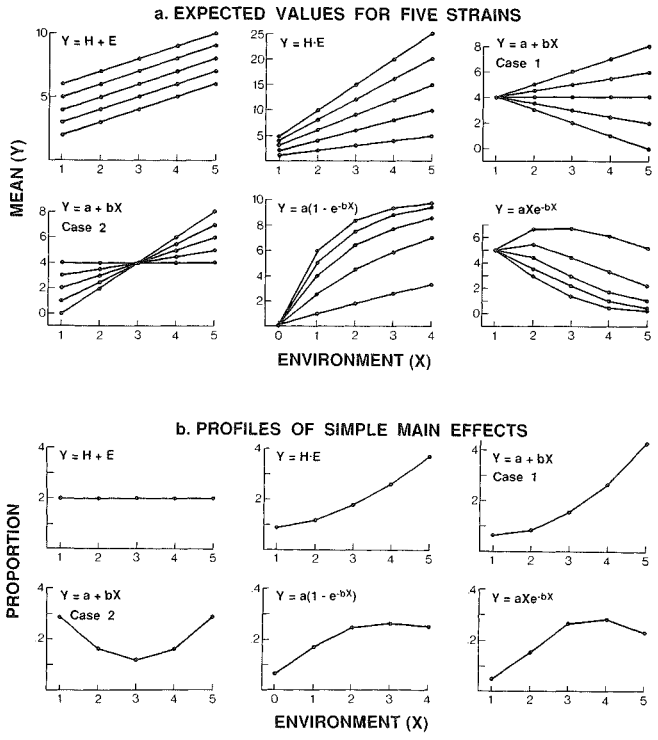


Figure 2. (a) Expected values of a measure Y under six models for five strains of mice reared in five different environments where levels of environment (X) are 1.0 units apart. Parameters of each model are assumed to be determined by each strain's heredity. (b) Profiles of simple main effects of heredity at each level of environment for the six models in Figure 2a, expressed as a proportion of the combined sum of squares for heredity and heredity by environment interaction.

The inescapable conclusion is that the usual application of two-way ANOVA is relatively insensitive to the presence of real nonadditivity of the kind considered plausible by many investigators. There are basically two views about this reality. If the principal objective is to partition variance and calculate heritability coefficients, this may be seen as evidence that analysis of variance is "robust" with respect to the assumption of additivity. On the other hand, if the goal is to understand the nature of development, the way things work, there will tend to be skepticism about a statistical procedure which takes data that, to the educated eye, show obvious differences in slopes and shapes of the norm of reaction for different strains, and apparently crunches them into a set of parallel straight lines. From the latter perspective, it will be

Table 4. *Effect sizes and power for six models using J = K = 5, n = 10 and α = 0.05*

	Effect sizes			Power of tests of		
Model	f_H	f_E	f_I	H	E	HxE
Y = H + E	0.40	0.40	0.00	>99	>99	—
Y = H·E	0.40	0.40	0.19	>99	>99	36
Y = a + bX, Case 1	0.40	0.00	0.28	>99	—	78
Y = a + bX, Case 2	0.00	0.40	0.28	—	>99	78
$Y = a(1 - e^{-bX})$	0.26	0.40	0.14	90	>99	19
$Y = aXe^{-bX}$	0.40	0.34	0.21	>99	>99	47

difficult to understand how any inquiry could possibly benefit from a test with inherently low power which often yields deceptively simple results.

It is noteworthy that techniques which estimate heritability by assuming no H × E interaction can yield strongly biased estimates when certain kinds of interaction are indeed present in the data. After a detailed mathematical study of path analysis, Lathrope et al. (1984) concluded that among "the principal effects of interaction are a mean overestimate of the genetic heritability" (p. 618). One hopes that this finding will not discourage investigators from seeking better ways to detect H × E interaction and its consequences.

9. The perception of simple main effects

If ANOVA is so insensitive to interaction, then alternative approaches are required. Taking the relationships in Figure 2a, which usually yield nonsignificant H × E interaction terms, let us construct for each one a profile of expected simple main effects of heredity or strain difference at each level of environment (Figure 2b). For each of these examples there are five levels of E and hence five simple main effects of heredity. The total of the sum of squares (SS) for these five must be equal to the SS for the main effect of heredity plus the SS for H × E interaction (Winer 1971). Thus, we can compute, at each level of E, the proportion of $(SS_H + SS_{HE})$ which is accounted for by that particular simple main effect. If the relation between H and E is truly additive, then that proportion should be the same across all levels of E (Figure 2b). On the other hand, the profile of simple main effects is markedly uneven for the other cases in Figure 2b. An obvious departure of this profile from a horizontal line should alert us to the possible presence of nonadditivity in the relation between H and E, and thereby caution us against accepting a null hypothesis as true merely because we cannot conclusively prove it false. The need to interpret the pattern of results by careful inspection is emphasized by Bolles (1988), whose Rule 5 is: "Always, always plot up the data to see what the numbers say. The numbers that you collect in an experiment will tell you if you have found something, even while statistical tests are fibbing, lying, and deceiving you" (p. 83).

The utility of this approach is sometimes recognized implicitly when scientists look at graphs from a two-factor experiment, perceive what appears to be interaction, and then do separate *t* tests to confirm this impression. But does this approach truly prove the existence of nonadditivity? One example suggests caution. The most widely cited report of heredity-environment interaction in psychology is the Cooper and Zubek (1958) study of the McGill "bright" and "dull" rat strains (bred selectively for errors on the Hebb-Williams mazes) reared in three laboratory environments (restricted, normal, and enhanced). (As Platt & Sanislow [1988] point out, the data for rats in the "normal" environment actually came from a separate experiment performed earlier.) The authors compared various pairs of the six groups, totalling only 65 rats or 10.8 per group, using separate *t* tests, and it is generally believed that this demonstrated H × E interaction (Platt & Sanislow 1988). However, an ANOVA on the six groups using raw data kindly provided by R. M. Cooper reveals significant main effects of strain ($F = 4.98$, $p < 0.05$) and environment ($F = 14.33$, $p < 0.01$) but *no significant interaction* ($F = 3.07$, $p > 0.05$). The *F* ratio for H × E interaction is slightly below the critical value of 3.15. Properly speaking, the data provide suggestive evidence but not conclusive proof of H × E interaction. The mere observation that two strains differ significantly at $\alpha = 0.05$ in one environment but not in another does not necessarily warrant rejecting the hypothesis of additivity. After all, one value of *t* might be just great enough to achieve significance while the other *t* falls a bit short of significance. In the Cooper and Zubek (1958) data, the strain difference was obviously large in one environment and small in the others, which was quite sufficient to convince most of us that there was H × E interaction.

10. Sample sizes for detecting interaction

It would seem that many studies of heredity and environment end up in a twilight zone of inconclusive results where different people can easily interpret subtle patterns in the data to be hints of this or that, where the fading hopes of some are kept alive by "almost significant" interaction effects or results "tending in the direction of significance," while others are relieved that the interaction effect was not quite large enough to rule out heritability calculations. From a statistical standpoint, the studies often lack sufficient power to shed much light on the nature of H × E interaction.

Looking closely at the data may help us avoid such serious mistakes as accepting a false null hypothesis, but the gaze of an experienced investigator is also fallible and can never be a complete substitute for a statistical test. Outright rejection of additivity really ought to require a significant interaction term in the ANOVA. If we are careful to avoid Type I errors when testing for the presence of main effects, surely we should also try to avoid them when testing for interactions. Why opt for a more complex model if it really isn't necessary?

Perhaps we would be wise to anticipate these various difficulties and address them at the design phase before data are collected. If the effect size for a plausible kind of interaction is substantially less than for the main effects, then a larger sample size will be required to detect the interaction than will be needed merely to detect average effects of the treatments. If it really matters whether or not the phenomena being studied are nonadditive, one needs to use larger samples than are customary for finding main effects. Proving nonadditivity false requires, at the very least, that the power of a test of interaction be substantial, 80% or preferably 90%, and that a proper sample size be chosen to guarantee sufficient power. Cohen (1977) provides convenient tables of sample sizes which yield different degrees of power for various effect sizes. A normal approximation that is useful when the interaction term has one degree of freedom is provided by Lachenbruch (1988). As shown above, for a two strain by two environment experiment the effect size for interaction under the $Y = H \cdot E$ model will be $f_I = 0.133$ when main effect sizes are large (0.40). The required sample sizes to detect such an interaction at powers of 80% and 90% with $\alpha = 0.05$ are more than 125 and 167 subjects per group, respectively, according to Cohen's tables. These values may appear extremely large, but the analysis of

variance with its definition of interaction as leftovers demands large samples. What reason could one possibly cite for using an analytical device because of its ability to detect nonadditivity, yet choosing a sample size that renders it ineffective? The finest optics in the world will portray a fuzzy image if the camera is out of focus or shaking.

11. Perils of ad hoc scale transformations

It is sometimes proposed that interaction in any kind of factorial design be addressed by transforming the scale of measurement to make the main effects additive. For example, Dunn and Clark (1974) recommend the procedure of Tukey (1957) whereby a computer is used to find the values of the constants C and p in the transformation $(Y + C)^p$ which minimize the size of the interaction term relative to main effects. In biometrical genetics in particular, the investigator is advised to search for a transformation that will eliminate heredity-environment interaction entirely, so that heritability and other parameters can then be estimated (Jinks & Broadhurst 1974, p. 11; Mather & Jinks 1982, p. 64).

Is this approach legitimate? Perhaps it is, if there is no other way to meet the assumptions of equality of within-group variances, normality, and independence of errors. When a mean-variance correlation occurs for response time measures or when many observations in some groups occur near the upper or lower limit of the scale, a transformation may be necessary to permit a valid test of significance, and such a transformation may also eliminate a two-way interaction. If the interaction does have a rather trivial origin in mean-variance correlation, then the transformation may be warranted. Even then, there may be pitfalls inherent in the procedure, because parameter estimates of the logarithm of a variable, for example, can produce biased estimates of the untransformed measure and can distort the estimates of variance components (Heth et al., 1989; Kvålseth 1985).

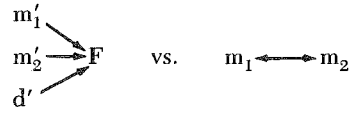
There has been some dispute in the pages of the *Psychological Bulletin* about whether the scale of measurement affects decisions about statistical significance, with arguments that it does not (Davison & Sharma 1988; Gaito 1980) and counterexamples showing that it can (Townsend & Ashby 1984), but this particular dispute has been focussed on comparisons of two independent groups. Concerning the consequences of transformation for two-way interaction, there is no doubt that conclusions can be drastically altered. The question is: *should* they be altered?

The model on which ANOVA is based assumes equality, normality, and independence of within-group deviations, but it does not assume additivity of effects, although path analysis does (Wright 1921). Transformation solely to eliminate interaction is a device to create the appearance of simplicity in the data, and there is a danger that this will be an entirely false appearance. For those who wish to learn how development actually works, wholesale and ad hoc testing of various transformations for the express purpose of getting rid of H × E interaction is counterproductive, because the shape of a functional relationship between variables provides a valuable clue to their causal connections. On the other hand, those whose only goal is to parcel out the variance among separate causes can proceed only in the absence of H × E interaction and therefore they may be more willing to transform the scale of measurement, even if causal relations become distorted.

To return to the gravitation example, we can see that transformation of scale can radically alter the causal or explanatory model. If we apply a logarithmic transformation to Newton's law, the equation becomes additive.

$$\ln F = \ln\left[\frac{Gm_1m_2}{d^2}\right] = \ln G + \ln m_1 + \ln m_2 - 2 \ln d$$

Physicists use this approach to analyze sources of measurement error, but they do so from a perspective very different from that of investigators who choose a transformation without knowledge of the form of a genuine law of nature. If we let the log transformed variables in Newton's law be the primed (′) variables, it reads: $F' = G' + m_1' + m_2' - 2\,d'$. The interpretation of this equation is altogether different from the real law if we forget about the transformation and take the terms at face value. Additivity implies a causal model which separates the contributions of the two masses, whereas the multiplicative



model implies mutual interdependence. Newton achieved a profound insight, which had eluded most predecessors who regarded the weight of an object as an inherent property of that object itself, something that existed in isolation from its surroundings. He argued that every speck of matter in the universe has mutual attraction with every other speck. Mutual attraction is expressed as the *product* of the masses. The weight of an object is the result of its interaction with other objects. It makes no sense to say that a person's weight depends more on body size than planet of residence. The additive model is really no simpler than the multiplicative one, in that both have three variables and a constant. The log transform alters the relations among the variables; consequently, transforming the scale of measurement may conceal the relations among heredity and environment, as it might conceal the essence of gravitation.

Transformation to suppress H × E interaction may create further obstacles to applying the knowledge gained from ANOVA. Consider the first use of ANOVA for a two-way factorial design by Fisher and Mackenzie (1923) to examine the yield of 12 potato varieties under six conditions of manure at the Rothamsted Experimental Station. Yields ranged from 26.5 lbs. per row for the Up to Date variety with farmyard dung to 1.6 lbs. per row for the hapless "undunged" Duke of York. The effect of sulfate of potash appeared to depend strongly on variety of plant and presence of dung, but the interaction term was not significant, although main effects were large. Inspecting their data, Fisher and Mackenzie observed a nonadditive pattern whereby higher yielding plants benefitted more from manure; they accordingly wrote that: "A far more natural assumption is that the yield should be the product of two factors, one depending on the variety and the other on the manure" (pp. 316–17). Rather than transforming their original observations of yields, they showed that the data "are better fitted by a product formula than by a sum

formula" (p. 320). Modern quantitative behavioral genetics, however, would dictate a transformation to achieve additivity. Such a procedure may be convenient for the theorist, but the everyday men of the soil must sell their potatoes by the pound and purchase manure by the ton. If there is a variety whose yield increases more than others for the same bulk of fertilizer applied, they would certainly want to know about this. After all, they cannot pay their bills in the square root of pounds sterling. To the farmer or scientist struggling to understand how things grow or develop, real interactions should not be hidden by ad hoc scale transformations.

Of course, transformation of scale need not conceal information. If we can discover a transformation that effectively eliminates interaction from the ANOVA, this reveals something about the mathematical structure of the original observations (Lubin 1961). There is no serious harm in generating additivity with a logarithm, provided the investigator remembers to calculate and report the anti-log when interpreting the results, rather than reifying the additivity. For example, Box and Cox (1964) used a log transform of a measure of strength of worsted yarn in a three-way ANOVA to demonstrate that the relations among weight of the load, length of yarn, and duration of loading are multiplicative because the log of strength eliminates the interactions. A problem arises when the original data are transformed and the profound effects of the change of scale on the causal model are neglected when presenting the results. If H and E really are multiplicative in a particular situation, a calculated "heritability" is nonsensical and taking the log of the observations may compound this.

12. Other approaches

The primary remedy proposed here for the problem of the low power of tests of interaction is the same as the one suggested by Neyman in 1935: Use larger samples, supplemented by a large dose of caution and rigor when interpreting results. Are other, possibly more palatable solutions available?

Neyman (cited by Traxler 1976) also proposed that additivity should be affirmed only if the main effects are significant at the 0.01 level, whereas interactions are not significant at the 0.05 level. Using different α levels could indeed reduce or even eliminate the imbalance in the power of the tests, although this could become rather cumbersome because the values of α required to equate the powers would depend on the specific alternative model being contrasted with the additive model. Furthermore, if the α level is set at 0.05 for the interaction term, the larger samples documented in section 10 are still required.

The interaction term in a J × K factorial design provides a global test of all possible kinds of deviations from strict additivity and hence may not be very sensitive to particular kinds of nonadditivity. It is possible to test more specifically for linear interactions whereby groups at different levels of one factor have different slopes of linear response to levels of the other factor or when the factors are thought to be multiplicative (Freeman 1973; Mandel 1961). Perkins and Jinks (1973) used a similar approach to show that large variations among 82 strains of tobacco plants in response to 16 fertilizers were almost entirely due to interactions of the linear type. These procedures will probably have greater power than the global F test, although the amount of gain has not been evaluated. However, there is concern that these tests may be biased (Roux 1984). Of course, they can provide no improvement at all for a 2 × 2 design and are needlessly complex for modest experiments having few degrees of freedom for the interaction term, which can be assessed more readily with orthogonal contrasts (Lachenbruch 1988).

A more radical departure from the standard ANOVA procedure is provided by the likelihood ratio test (Marler 1980), which compares the likelihoods of a particular set of data according to two distinct hypotheses, neither of which must serve by default as the null hypothesis. This approach can be extended to more than two reasonable alternatives, as done by Debray et al. (1979). Similarly, one could compare additive and various nonadditive models of heredity and environment in a factorial design. These calculations require much more effort from the investigator and considerable computing time, but they should yield greater statistical power than the ANOVA approach. Unfortunately, this would also require much greater mathematical knowledge on the part of the reader.

13. Heritability and eugenics

Analysis of variance may be useful in identifying significant sources of individual differences, but its insensitivity to the underlying mathematical structure of functional relationships limits its utility to the early phases of investigation. If variations in both heredity and environment are found to contribute to individual differences in behavior, then the next phase of the research ought to look more closely at the intricacies of the two processes in the developing organism using larger samples and more sensitive analytical methods. Simply to cite a heritability coefficient or compare the relative strengths of the main effects of heredity and environment in a factorial experiment does not advance our understanding of the nature of development.

Unfortunately, estimating heritability seems to be the main objective of some investigators. As Kevles (1985) and Fancher (1985) have documented, many of the founders of human behavioral genetics were committed to a program of eugenics. The only practical application of a heritability coefficient is to predict the results of a program of selective breeding. The rate of change in the average value of a characteristic during the first few generations under a regime of artificial selection of breeders will be directly proportional to the heritability (in the narrow sense) of the characteristic in the population. If such a goal is eschewed, there is no compelling reason to focus attention on "heritability" and ignore interaction.

14. Gene action is interactive and dynamic

Of course, statistical problems are not the only challenges to theories of additivity of heredity and environment, and statistical solutions are not likely to settle this dispute. Perhaps the greatest weakness in the Y = H + E model is the assertion that the effects of one's heredity on develop-

ment are entirely separate from those of one's environment. This claim is contradicted by many discoveries in developmental biology.

There are now good reasons to believe that the genes in the nucleus do not contain a program for development or a blueprint for brain structure (Gerhart 1982; Stent 1981). The timing and spatial location of important events in development are not directly specified by information intrinsic to the genes in the nucleus (Davidson 1987; Easter et al. 1985; Oyama 1985). Rather, a gene codes or programs for a protein or enzyme, and the consequences of this activity at the level of macromolecules for events at the cellular and organismic levels, depend on other parts of the cell, other cells in the growing organism, and even events outside the organism. The metabolic activities of DNA molecules are subject to control by factors outside the nucleus of the cell (Blau et al. 1985). The actions of certain genes can be modified greatly, even sometimes switched on or off entirely, by changes in temperature (Atkinson & Walden 1985; Heikkila et al. 1986), light (Klein & Yuwiler 1973), diet (Benkel & Hickey 1987), and even the maternal environment (Carroll et al. 1986). Developmental biology is tuned in to nonadditive processes (Pritchard 1986). Direct evidence of biochemical gene action in an environmental context supports a dynamic and interactive view.

The continued use of statistical tests insensitive to interaction is distressing, not merely because it fosters a false impression that heritability analysis is justified, but because valuable information about processes of development may be lost. A knowledge of interaction deepens our understanding of how living things acquire form and motion. According to Lubin (1961): "The most important questions that can arise from a statistical finding of interaction are those which are non-statistical. . . . For me, significant interactions raise two most important questions: How does this interaction occur? How can I bring it under experimental control?" (p. 816). Likewise, for Lassalle (1986), H × E interactions should be viewed "as powerful tools which can assist us in understanding the underlying processes of behaviour" (p. 205), and for Bateson (1987) "analyses of statistical interaction should be the starting points of attempts to understand how developmental processes work and should not be treated as ends in themselves" (p. 2).

ACKNOWLEDGMENTS

Supported in part by grant 4878 from the Natural Sciences and Engineering Research Council of Canada. I thank Margot Andison for drawing Figure 1 and Susan van Ballegooie for typing the manuscript. The initial version of this paper was written when the author was at the Department of Psychology, University of Waterloo, Ontario.

NOTES

1. Effect size f for a one-way ANOVA is related to an alternative measure of effect size, the proportion of total variance attributable to differences among group means, termed η^2 by Cohen (1977) and ω^2 by Hays (1988), according to the relation

$$\omega^2 = \frac{f^2}{1 + f^2}$$

For ω^2, small, medium, and large effect sizes would be about 0.01, 0.06, and 0.14, respectively. Cohen (1977) gives power in terms of f, but several other sources use the noncentrality parameter of the noncentral F distribution, λ, or related measures δ or ϕ, with the following relations among them for J groups of n observations each:

$$\phi = f\sqrt{n} \quad ,$$
$$\delta = \phi\sqrt{J} = f\sqrt{nJ} \quad ,$$
$$\lambda = \delta^2 = f^2nJ.$$

2. Principal steps in the derivation of σ_E are

$$M = \frac{(J + 1)(K + 1)he}{4} \quad ,$$

$$M_k = \frac{k(J + 1)he}{2} \quad ,$$

$$M_k - M = \frac{(J + 1)he}{2}\left[k - \frac{(K + 1)}{2}\right] \quad ,$$

$$\sum_{k=1}^{K} (M_k - M)^2 = \frac{(J + 1)^2(he)^2K(K + 1)(K - 1)}{48} \quad .$$

For finding σ_I, the first step for each group is:

$$M_{jk} - M_j - M_k + M = \left[j - \frac{(J + 1)}{2}\right]\left[k - \frac{(K + 1)}{2}\right]he.$$

Across all J·K groups, this yields:

$$\sum_{j=1}^{J}\sum_{k=1}^{K} (M_{jk} - M_j - M_k + M)^2 = \frac{JK(J+1)(J-1)(K+1)(K-1)(he)^2}{144}$$

3. The tables in Cohen (1977) and most other sources on power of ANOVA apply directly to a one-way design, but our interest here is in a two-way factorial design. Cohen (1977) addresses this problem by noting that a mean for one level of the first factor across all levels of the other is not based on only n observations; rather, it is based on nK observations. Of course, a few degrees of freedom are lost because of constraints placed on the data in computing between-groups sums of squares; hence, the effective sample size (n′) for a test of the main effect of heredity is

$$n' = \frac{df_{error}}{df_{between} + 1} + 1 = K(n - 1) + 1.$$

When there are 5 strains reared in 5 environments and $n = 10$ subjects per group, effective sample size per strain for the test of the main effect is 46. For the test of interaction, $n' = 14.2$ because

$$n' = \frac{df_{error}}{df_{HxE} + 1} + 1 = \frac{JK(n - 1)}{(J - 1)(K - 1) + 1} + 1.$$

That is, the power of the test of the interaction term is essentially the same as the power of a test of variation among 17 groups with 14.2 observations per group in a one-way design.

Rather than deriving all values of power by interpolation from the tables given by Cohen (1977), the normal approximation to the noncentral F distribution (Severo & Zelen 1960) was used. This is not the best available approximation (Tiku 1966), but it is reasonably good when we are interested in statistical power to only two decimal places or the nearest percent, and it is much easier to compute.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

An interaction effect is not a measurement

Fred L. Bookstein

Center for Human Growth, Department of Biology, University of Michigan, Ann Arbor, MI 48109-0406
Electronic mail: *fred_l._bookstein@um.cc.umich.edu*

Wahlsten's target article inquires about when it "makes sense to attribute a definite percentage of variation in some measure of behavior to variation in heredity." Yet the discussion does not touch upon technical aspects of heritability per se, such as the relations between parental and offspring scores. Instead, the fallacies of the behavioral genetics literature surveyed here exemplify confusions about the meaning of quantification in science that run far deeper.

To ask, with the target article, "$Y = H + E$ – true or false?" is to conceal the underlying issues of mensuration. The scientist and the user of analysis of variance (ANOVA) approach the issue of describing Y, H, and E radically differently. Whereas the scientist would demand that E and H be measured in their own units, the user of ANOVA implicitly "measures" categories of E and H using the units of Y instead. When main effects are measured in units of the dependent variable, then the $H \times E$ interaction has *no operational meaning* as an amount of anything "per" anything else; it has no units; it is not a measurement. The relevant aspect of Newton's law of gravitation, reviewed in the target article, is that masses are measured in units of mass, not in units of force. It is this possibility of direct measurement, regardless of functional form, that saves the physicist from the futilities of ANOVA.

Once H and E are measured directly, the issue of "interaction" is trivial. Temporarily, in the interest of clarity, assume that we already know how to measure H and E as numerical variables x and z, and that the true law describing the dependence of Y on x and z is some noise-free function $Y = f(x,z)$. In this setting, the condition equivalent to absence of the interaction term in an ANOVA of Y upon H and E is that f be "separable":

$$f(x,z) = g(x) + h(z), \tag{1a}$$

the mathematical functions g and h replacing the familiar "main effects" of H and E. The assertion of separability is a single partial differential equation of second order:

$$\frac{\partial^2 f}{\partial x \partial z} = 0 \text{ everywhere.} \tag{1b}$$

But no particular interest inheres in this separability condition in the absence of other equations governing the behavior of g and h. (For instance, in classic studies of intergenerational heredity, g takes the form of a linear regression to the mean.) When such additional equations are not forthcoming, the knowledge of separability (absence of interaction) is of no particular value, as it can always be guaranteed by small changes in the measurement system. Define a *critical point* of f as an (x,z) pair where both first-order partial derivatives

$$\delta f/\delta x \text{ and } \delta f/\delta z$$

are zero *at the same time*. It is a theorem that any $f(x,z)$ without critical points in a region of interest can be expressed as perfectly linear in some deformed version of the (x,z) – coordinate system. The theorem, a prologue to the so-called "Morse Lemma" (Poston & Stewart 1978, sec. 4.2), guarantees that we can write

$$f(x,z) = f_0 + \left(\frac{\partial f}{\partial x}\right)_0 x' + \left(\frac{\partial f}{\partial z}\right)_0 z' \tag{2}$$

(no further terms), where $x' = x +$ small stuff and $z' = z +$ small stuff are gently nonlinear functions of the originally measured x and z. For an exploration of this and other aspects of catastrophe theory, refer to Poston and Stewart.

Thus, in the absence of critical points we can always make an interaction term disappear by slightly altering our scheme for measuring x and z. The presence of critical points is far more important than this: It is *generic*, not an accident of small changes in measurement systems. Consider (Figure 1) the two surfaces $y = -x^2 - z^2$ (left) and $y = x^2 - z^2$ (right). Neither has any "interaction terms" like xz – but they could not be more different in their scientific implications. Whereas the former has a global maximum at one point (0,0) of the domain of its arguments, the latter goes off to $\pm\infty$ twice each as one goes around lines through the origin. (The cases are distinguished by the sign of the *discriminant* $B^2 - AC$ of the quadratic approximation $Ax^2 + 2Bxz + Cz^2$ to Y.) Rewriting $x^2 - z^2$ as $(x + z)(x - z)$, we see again that, as the target article avers, there is no intrinsic difference between additive ("interaction-free") and multiplicative models. Instead, there remain huge differences among varieties of critical points *even in the absence of interactions*. The presence of a "statistically significant" interaction term – say, $2xz$ – doesn't tell us whether the situation is free of critical points or, if not, whether it resembles more the extremum, the saddle surface, or some other more exotic possibility.

In a stochastic setting (noise present), the separability condition (1*a*) or (1*b*) is replaced by a statement about expected values: for any z_1 and z_2, we are asserting that $E[f(x,z_1) - f(x,z_2)]$ is independent of x (from which it follows that $E[f(x_1,z) - f(x_2,z)]$ must be independent of z). In most applications, $E[f(\cdot,z)]$ will then exist as an expectation, for fixed z, over the universe of possible x's as they substitute for the dot under stratified sampling, and likewise $E[f(x,\cdot)]$ exists as an expectation for fixed x over stratified sampling of zs. But, as we have seen, it is not the existence of these expectations that is the scientific issue here – we can almost always redefine x and z jointly so that these expectations exist and can be computed either via formula (2) or by sums of squares analogous to those in Figure 1. It is the *forms* of the functions $E[f(\cdot,z)]$ and $E[f(x,\cdot)]$ that are crucial, specifically, the number and genres of their critical points, regardless of any interactions.

Analysis of Y, H, and E culminates in scientific understanding of the determination of Y only when the factors H and E have been explicitly measured separately and independently. The mensural fallacy underlying the literature criticized in the target article arises in the confusion between the H and E of "$Y = H + E$" – continuous variables x and z – and H and E as the mere *names* of "groups" or "strains" or "conditions" not otherwise calibrated. In the absence of interaction, the ANOVA computes $E[Y|H = h]$ and declares it to be the "value" of each state h of H,

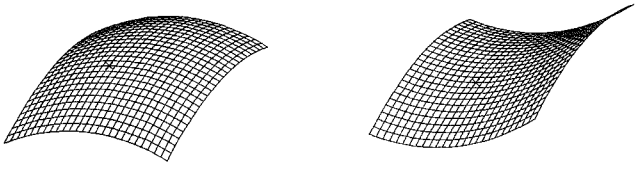


Figure 1. (Bookstein). Two mathematical surfaces, each free of interaction terms, which show essentially different behavior in the vicinity of their critical point. (left) $y = -x^2 - z^2$. (right) $y = x^2 - z^2$.

in units of Y. The interaction term is not a measurement; it is merely that single word of (risky) authorization.

The scientist needs to know the form of f; its separability can always be arranged a posteriori. Once this functional form is known, the difference between $Y = H + E$ and $Y = HE$ is intellectually and scientifically trivial; without knowledge of the form of f, path modeling or any other sort of causal modeling is actively misleading. It follows that it never makes sense to "attribute a definite percentage of variation in some measure of behavior to variation in heredity" at all, or to have a theory "rise or fall with the occurrence or absence of statistical interaction." Such procedures and decisions confuse the provenance of values of single measures with statements about the relationships among multiple measures. Once direct quantification of H and E is at hand, we are able to inspect the function f directly, and we thus have no need to partition any sums-of-squares involving Y. In measuring main effects *only* by their consequences, ANOVAs quantify ostensible causes in the wrong units, while the "interaction effect" is a measurement of nothing at all.

Methodological heterogeneity and the anachronistic status of ANOVA in psychology

Daniel Bullock

Cognitive & Neural Systems Program, Boston University, Boston, MA 02215.

Electronic mail: *danb@bucasb.bu.edu*

Wahlsten's instructive reminder of ANOVA's marked additivity bias forces us to confront the following question: What can be said of any research enterprise that appears to be unaware that its primary data analytic technique exhibits such a bias when used with typical sample sizes? Either (1) the research enterprise is of poor quality because it lacks the secondary consistency tests that would typically reveal the additivity bias, or (2) the secondary tests are in place but fail to reveal any method bias because the phenomena really are additive, or (3) the hypothetical "researchers" care more about sustaining the thesis of additivity than about gaining a more veridical understanding of the subject of research.

In the case of the natural subject matter for an unbiased science of behavior genetics, it is extremely unlikely that the effects of heredity and environment are for the most part truly additive. Thus if behavior geneticists have heretofore found the additive model near-universally applicable, as Wahlsten claims, then either they have failed to incorporate the kind of secondary checks essential for good science, or they have systematically ignored abundant counterindications. In either case, it would be reasonable to conclude that many past claims by behavior geneticists are unreliable.

Beyond that, it is important to diagnose the problem correctly and to improve the future record. Wahlsten focuses on ANOVA's additivity bias and recommends using sufficiently large numbers of subjects per cell to ensure that ANOVAs will have enough power to detect interactions. However, this strategy does not solve the root problem: Behavior genetics needs additional methods that afford consistency checks against results obtained with ANOVAs. Unlike other scientific enterprises, behavior genetics (as represented by Wahlsten) has allowed itself to become defined by its use of a single method: heritability analysis. This is rather like someone who claims to be a carpenter but refuses to use any tool other than a hammer: It is not surprising that such a carpenter professes to see nothing more in the world than lots of raised nails.

If Wahlsten's representation of the field is accurate, we should probably conclude that the phrase used above, "unbiased science of behavior genetics," is an oxymoron. No enterprise based on a single measurement device can be an unbiased science of anything. The solution is to add whatever methods behavior geneticists use to the broader toolkit of developmental psychobiology and cease talking of behavior genetics as a separate science. Though Wahlsten implicitly suggests that the result will be a net loss of interest in heritability analysis, it might instead be a transformation of the role of heritability analysis. Whereas it is of little scientific interest merely to partition phenotypic variance between heredity and environment, it would be of considerable scientific value to understand why some behavioral phenotypes fit an additive model whereas others do not.

More generally, though, interest should shift to characterizing the heterogeneity of types of heredity/environment interaction in behavioral development (e.g., Bullock 1987; Oyama 1985). Here progress will depend on proposing quantitative models of actual processes, and testing such models at the level of behavioral phenotypes as well as at any other levels, e.g., the neural networks level, known to be "interposed" between genes and behavior. In this context ANOVA will no longer serve as more than a minor player, because it will be possible to predict, hence tailor a test for, specific nonadditive effects.

In the search for types of interaction, it is useful to remember the basic logic of evolution. For Darwin, the primary observation was the correlation between characteristics of organisms and features of their local habitats. His and subsequent explanations of how this pattern arose made reference to some variants gaining a competitive advantage in a fixed environment and eventually filling the population in that specific environment with their descendents. This suggests that important residual variation in a population (the usual subject for behavior genetics studies) will often relate to the efficiency with which a lineage makes use of specific environmental resources. On the other hand, when environmental resources are plentiful, the more efficient organism may have no real competitive advantage because even inefficient organisms flourish. Finally, there are often limits to how much of a phenotype is useful to an organism: Too many fruits break the limb. These considerations suggest that one may often see a highly nonadditive pattern in which lines for different strains diverge from an initial point but reconverge at some high value of the phenotype as the environment becomes quite favorable relative to all strains' efficiencies. This pattern is akin to the fifth shown in Wahlsten's Figure 2, for which ANOVA has less detective power than for all the other models shown (see Table 4, fifth row).

The point about behavior genetics usually pertaining only to residual, individual difference variation, is a critical one. The basic covariation of species-general characteristics with average expected environment often means in practice that one can expect an inverted u-shaped function between an environmental quantity and phenotype value. For example, too much fertilizer or too much water kills the plant. By sampling only similar strains and only near the mean of the average expected environment, researchers may be able to find locally additive or near-additive relations, but this hardly gives an accurate picture of the system under study. It is particularly problematic in those species which, far from leaving the occurrence of an average expected environment to chance, actively construct and maintain such an environment. Of course, humans are the most extreme examples of such a species (Bullock 1987).

It might seem that human intelligence is a counterexample to the claim that most phenotypes are useful only up to a certain point. Intelligence is a special case because it is what may be called a second-order behavioral phenotype, that is, a phenotype that functions to abet the development of other behavioral phenotypes. It is also a very distributed phenotype, having many components that have emerged at different times in phylogeny. The ultimate effect of further increments in intelligence is to increase the rate of skill development. That is why it is best measured by an essentially open ended set of tasks: What

one really measures is differences in the speed with which testees have explored and internalized the space defined by the set of language games (Wittgenstein 1953), or modules of cultural practice (Adams & Bullock 1986), jointly available to tester and testee. I use *testee* rather than *individual* because the speed with which someone explores and internalizes modules of cultural practice depends heavily on teachers, who are a critical evolved component of the *intelligence* of a developing human (Bullock 1987). Needless to say, such a situation poses special problems for researchers whose goal is an additive partitioning of genetic and environmental sources of variance.

As Wahlsten recounts, ANOVA was invented by Fisher to serve in the assessment of agricultural yield experiments, not as a key tool for the developmental sciences, which seek to explain how structure emerges in nature. Equally to the point, ANOVA was invented before generalized regression analysis, before most of the modern field of measurement theory came into existence, and before high-speed computers removed computational complexity as a critical determinant of the practicality of using most data analytical procedures. It was also invented before it became feasible to study *n*-body, highly interactive dynamical systems via numerical simulations. Because of these interim developments, the special status of ANOVA in psychology – it remains the primary and in some cases the sole paradigm of applied mathematics in many psychology curricula – is an utter anachronism and extremely destructive of intellectual potential among young research psychologists. I agree completely with Wahlsten's implication that performing endless partitionings of variance has almost nothing to do with promoting scientific understanding in Piaget's sense, where "to understand is to reconstruct." But the remedy is to train many more psychologists as better system modelers and applied mathematicians, not solely to urge interpretive caution and larger samples, a recommendation that might even perpetuate the undue hegemony of ANOVA.

Interaction between genotype and environment: Yes, but who truly demonstrates this kind of interaction?

Michèle Carlier and Catherine Marchaland

Génétique Neurogénétique et Comportement, URA 1294 CNRS, UFR. Biomédicale, Université Paris V, 45 rue des Saints Pères, 75270 Paris Cedex 06, France

That Wahlsten's target article called for a commentary from two specialists of two different disciplines is proof of its interdisciplinary value.

From the genetic standpoint. Finding an interaction between genotype and environment (GxE) helps to identify some of the physiological pathways from genes to behavior. For example, the effect of a given treatment could be considered a compensation for possible defects of given genes. Roubertoux (1981), Carlier et al. (1983), and Carlier & Robertoux (1986) developed this point of view before others who are cited by Wahlsten. One can wonder, however, when experimental design makes it truly possible to detect a GxE? Let us take the example chosen by Wahlsten (section 4.1) which is a very common design in behavior genetic research:

J strains are reared in *K* different environments. The two independent variables are the strain and the environment. Thus the interaction, if interaction there is, is a strain *x* environment interaction and not a *genotype x* environment interaction. In fact, different strains (inbred or selected) differ both in genotype and in maternal environments (cytoplasmic, uterine, and postnatal) provided to offspring. Thus the inference from the strain effect to the genotypic effect is based on the assumption that these maternal effects are null. This mistake is very easy to make. One of the present commentators did so herself in a paper on the GxE (Carlier & Roubertoux 1986, p. 71) when presenting results observed on pup development in mice. Three years earlier she had pointed out the risk of this kind of error in a paper. In the title of this very paper she herself made the mistake (Carlier et al. 1983). This mistake is of great significance in human behavior genetics, however. Authors come to interpret as a GxE effect data collected using a cross-fostering design where the independent variables are the characteristics of the biological parents and the adoptive parents (e.g., IQ, economic status, etc.).

In a recent paper Roubertoux et al., (in press) noted the scarcity of experimental papers demonstrating a true GxE. This might be due to the high technology required to show such an interaction (Carlier & Nosten 1987; Carlier et al., in press). In this field, without a straightforward separation between the genotypic effect and the different components of maternal environment (cytoplasmic, uterine, and postnatal), nothing can be demonstrated. Ovarian transplantation (or egg transfer) jointly implemented with fostering provide the means of separating these components. To our knowledge, Nosten & Roubertoux (1988) have presented the first true GxE in two inbred strains of mice, CBA/H (H) and NZB (N). An effect of parental versus F1 uterine environment on age at eyelid opening appeared in H pups and not in N pups. Nosten (1989) presents other evidence for a truly interactive effect on early developmental markers and these effects are not limited to the preweaning period (Roubertoux et al., in press).

From the statistical standpoint. Three points have to be stated once again: (1) an effect is more or less established according to the level of significance; (2) the level of significance of the *F* test only provides information about the existence of a theoretical effect but provides no information about the effect size: (3) the nonsignificance of the *F* is no more than an admission of ignorance and one cannot conclude from it that there is no effect. We are not proposing here an alternative to the ANOVA. However, we are proposing the use of Bayesian methods as a cure for some of its shortcomings.

The Bayesian methods were developed by Rouanet et al. (1978), then improved by Rouanet & Lecoutre (1983). As in using ANOVA, the distribution of the data is considered normal and the variances are considered homogeneous. Bayesian procedures give information about the magnitude of the effects. Let us consider d the observed effect and δ^* the theoretical effect. From the d and the F ratio one can establish a probability distribution which takes the same form as a generalized Student t with center d and scale $e = |d| / \sqrt{F}$ (with q d.f.). If q is high the Bayes-fiducial distribution is approximately normal with center d and scale e. With the Bayes-fiducial guarantee fixed at γ one can infer a statement which gives decision rules about the size of the effect δ.

If F is significant, a "noticeable effect" is tested with:

$$P(\delta^* > ?) = \gamma \quad \text{si } d > 0$$
$$P(\delta^* < ?) = \gamma \quad \text{si } d < 0$$

If F is not significant, a "negligible effect" is tested with:

$$P(|\delta^*| < ?) = \gamma$$

It must be pointed out that sometimes no decision can be made. For example, when testing an interaction effect, if the F is not significant and if the negligible effect does not appear, it is not possible to conclude anything without the risk of committing a type 2 error.

An alternative solution suggested by Wahlsten is to duplicate the experiment with higher levels of freedom. Would it be possible to be less demanding for the significance level of the interaction?

These Bayes-fiducial methods do not answer all of Wahlsten's questions because they use the F test, which is the target of

Wahlsten's critique. They demonstrate that all is not lost, however, when the F is not significant.

ACKNOWLEDGMENT

Preparation of this commentary was supported by the CNRS (URA 1294), the MEN (University Paris V and University Reims Champagne Ardenne), and the Foundation Pour la Recherche Médicale.

Inheritance and the additive genetic model

James M. Cheverud

Department of Anatomy & Neurobiology, Washington University School of Medicine, St. Louis, MO 63110

Wahlsten emphasizes the importance of genotype by environment interaction effects in the causation of behavioral phenotypes and points out that the usual ANOVA model gives primacy to additive effects, thus disguising the presence and relative importance of interactions in the developing phenotype. I believe that this problem with heritability analysis occurs because researchers attempt to use a model developed for the study of inheritance to investigate the developmental mechanisms underlying phenotypes. Quantitative genetics is concerned with the statistical effects of genes on phenotypes, not their physical, physiological, or mechanical effects or actions. Gene effects should not be reified as gene actions. These statistical effects are partitioned so that heritable and nonheritable variance can be separated from one another. They are not partitioned in a manner which leads to necessarily important distinctions about the physiological development of a character.

Quantitative genetics developed out of the search for a theory of inheritance which was consistent with Mendel's laws, but could account for the inheritance of complex characters (Provine 1971). An accurate theory of inheritance was needed to formalize Darwin's theory of evolution. Fisher (1918) succeeded in doing so and developed the basic ANOVA model in order to derive the covariance among relatives based on Mendel's laws, thus providing a quantitative theory of inheritance for complex traits. According to this theory, only the additive effects of genes can be passed on from parents to their offspring because, in a diploid population, each parent contributes only a single allele to their offspring at any given locus. For this reason, dominance effects, which result from the interaction of the two alleles carried at a locus, and the variance associated with them, cannot be inherited in a random-bred population. The additive or breeding value of an allele is the crucial characteristic in a model of inheritance. However, it does not depend only on the action of a gene, but also on the frequency that gene has in the population (Falconer 1981).

The concepts of additive effect and heritability – the proportion of phenotypic variation which is inherited – play very important roles in evolutionary models and empirical studies of inheritance which are directed towards evolutionary problems (Falconer 1981; Lande 1976; 1979). The concepts were developed for this purpose and are currently accepted as successfully applied in an evolutionary framework. Since genotype by environment interactions do not bias heritability estimates, they are often not considered explicitly, although important evolutionary models including genotype by environment interaction have recently been developed (Via & Lande 1985).

It is not surprising that the additive genetic model, which was developed from and as a theory of inheritance, is perhaps less informative concerning gene actions or physiological effects. The distinction and relationship between the statistical and physiological effects of genes is an important one to keep in mind (Cheverud 1984; 1988). Quantitative genetics is concerned with the statistical effects of genes since these effects are important for evolution. Genes also affect development in a mechanical or physical sense. When these gene actions have phenotypic consequences and vary across the population, they will have statistical effects. Genes act in development, regardless of whether they vary in a particular population. Thus the additive genetic model may not be suited to Wahlsten's goal, which is to "understand the nature of development" (sect. 8, para. 3).

If heritability analysis is not precisely suited to the study of development and gene action, why is it so commonly used and what can replace it? First, I believe that it is commonly used because it can at least provide evidence that genes play a role in the development of a particular phenotype. We can only detect gene action through statistical variation in gene effects. Molecular genetics is now beginning to allow direct experiments on gene action, but this work is still in its infancy, especially with regard to complex morphological and behavioral phenotypes. Thus, a significant main effect in an ANOVA should only be taken as evidence that genes affect individual differences in a phenotype. The fact that the effect is additive is important only for analyses of inheritance and evolution, not for development. The further test for genotype by environment interaction is performed as described by Wahlsten (1989) and is appropriately conservative. We should only consider the possibility of this interaction after having first shown that genetic variation causes variation in phenotypes. Again, this should not be considered as a good estimate of the relative degree of interaction, but rather only as evidence that interaction does occur. It is an appropriate test for this circumstance since interaction is a statistically and physiologically more complex developmental model than separate additive environmental and genetic effects.

The only way to improve the genetic analysis of development would be to generate developmental models which would help guide the analysis, just as the Mendelian model of inheritance guided the development of ANOVA. Riska (1986) and Slatkin (1987) have presented developmental models and investigated their consequences for patterns of heritable variation. Wahlsten simulates several nonadditive developmental models and shows that ANOVA does a poor job of statistically detecting the physiological interaction simulated. If an accepted model of phenotype development could be generated, the form of statistical analysis would be dictated by that model, just as the form of ANOVA was dictated by the Mendelian model of inheritance. Just as no one would analyze gravitation with an additive model, knowing the multiplicative relations between force of attraction, the masses of the objects, and their separation, no one would analyze development with an additive model if some reasonable quantitative model of development were available. The problem is not with ANOVA, but with the lack of developmental theory.

Huxley's (1932) model of relative growth is a good example of a developmental model which drives the statistical analysis of data. Because of Huxley's model of multiplicative growth, most studies of allometry apply linear models only after logarithmic transformation. This transformation is motivated by practical considerations concerning ease of calculation, since the multiplicative model could be estimated directly with maximum likelihood methods.

ACKNOWLEDGMENT

This research was supported by NIH grant 7 RO1 NS24904.

Additivity, interaction, and developmental good sense

David A. Chiszar and Eugene S. Gollin

Department of Psychology, CB 345, University of Colorado, Boulder, CO 80309

A popular text on evolutionary biology contains the following passage:

Very often the reaction to a difference in environment differs

among genotypes. There is then a Genotype × Environment Interaction ($V_{G\times E}$) that contributes to the phenotypic variance, so that $V_P = V_G + V_E + V_{G\times E}$. In this case the genetic variance is not readily distinguishable from the environmental variance because each depends on the other (Lewontin 1974; Gupta & Lewontin 1982). We will proceed, as most workers in the field do, by ignoring the Genotype × Environment Interaction, which in practice is often included in the term V_E. (Futuyma 1986, p. 197)

A major point in the target article states that ignoring significant $V_{G\times E}$ is a dubious practice because it makes estimates of heritability difficult, even impossible, to interpret. Another point raised by Wahlsten is that commonly used statistical tests of interaction do not have adequate power, suggesting that Type II error rates are distressingly high, not only in the assessment of $V_{G\times E}$ but also in other scientific domains where interactions are evaluated in the context of factorial experimentation and factorial ANOVA techniques. It is the juxtaposition of these two points that has startle value. The first one has long been clearly recognized (see references cited by Futuyma 1986), but it continued to be ignored because investigators had a false sense of confidence in the veracity of statistical tests that failed to confirm the presence of interactions. Wahlsten's articulation of the power issue now forces a reassessment of the tradition of ignoring $V_{G\times E}$.

Edwards (1985) points out "that in experimental work significant main effects are more common than significant two-factor interactions, and significant two-factor interactions are more common than significant three-factor interactions" (p. 242). Many reasons probably contribute to this state of affairs, including the power issue developed in the target article. To this argument we add another suggestion. Numerous statistics texts discuss the robustness of ANOVA vis à vis violations of the assumptions of normality and homogeneity of variance, usually citing Box (1954), Norton (1952), Boneau (1960), and Baker et al. (1966). These studies concentrate on single-factor experiments (hence, on main effects) and on Type I errors. Only a few studies have evaluated ANOVA robustness with respect to Type II errors, and these too have concentrated on main effects (e.g., Donaldson 1968; Tiku 1971). Hence, it is unclear that factorial ANOVAs, interaction factors in particular, are robust against departures from normality and homogeneity of variance. Furthermore, both Type I and II error rates need to be considered in Monte Carlo simulations designed to study these matters. Since violations of these assumptions are common in behavioral research, and since the demonstrable robustness of single-factor ANOVA has created a sense of complacency regarding such violations, it is possible that the inherently low power of typical ANOVA tests of interactions has been exacerbated by an incremented Type II error rate induced by violations of foundational assumptions regarding within-condition variances. The upshot of Wahlsten's insights plus this commentary is that Edwards's actuarial statement about interactions may be not only a comment on the structure of behavioral processes, it may also represent the historical accretion of Type II errors that have been generated for several different reasons. We hope that Monte Carlo or other appropriate techniques will be applied to test the suggestions made herein. The issues at stake extend far beyond behavioral genetics and evolutionary biology; indeed, they embrace the entire actuarial state of affairs captured by Edwards's summary remark.

Box (1953) made an analogy between tests of homogeneity of variance executed prior to ANOVA and putting to sea in a row boat to learn whether the water is safe for an ocean liner. In that context the remark was reasonable, yet considerations raised by the present exchange are on a more titanic scale.

The use of an inadequate statistical model, as is so clearly shown in the target article, is directly related to a faulty notion of development. It is based on a commonly held assumption, often implicit, by human behavioral geneticists, that hereditary influences can be made intelligible without recourse to consideration of extra-hereditary factors. Without such an assumption reliance on an additivity model is untenable.

A contrary view has been put forward by Alberch (1983), who has written that even if the complete DNA sequence of an organism were known, its morphology could not be reconstructed without knowledge of the epigenetic interactions that generate the phenotype. What is required is the adduction of the organizing principles that mediate development. Such a position seeks to account for development by emphasizing the in-principle inseparability of organism and environment (Chauvin 1977). It follows from this position that reductive explanations are insufficient to account for the complex patterning involved in psychobiological development (Gollin 1985).

Wahlsten is exactly on the mark when he emphasizes that the issue is not *heritability analysis* but rather the loss of information about developmental processes. Those experimental attempts to clarify the developmental process – for example, an organism-level X task-levels design – are thwarted by their reliance on a statistical device that raises severe barriers to the elucidation of interactive relationships. The primary objective of such experimental designs is the generation of a matrix of data that reveals both behavioral and biological *differences* that relate to organismic variation (e.g. age, strain, pathology, etc.), and behavioral and biological *similarities* that exist in spite of this variation. It is from the *pattern* of differences and similarities obtained from the Organism X Task analysis that an interpretation of the contribution of organismic variation to the functional system derives (Gollin 1965; Castro & Rudy 1989). The very insensitivity of the ANOVA design, as Wahlsten so clearly demonstrates, is what allows additivity models to be retained; it also encourages the avoidance or rejection of those developmental models which stress the inseparability and interdefinition of genome, organism, and ecosystem.

On the insensitivity of the ANOVA to interactions: Some suggested simulations

Domenic V. Cicchetti

VA Medical Center and Yale University, West Haven, CT 06516

Wahlsten presents cogent arguments to support his conclusion that the analysis of variance (ANOVA) is quite insensitive to heredity (H) environment (E) interactions. In developing his argument, the author discusses the important issues of power, effect sizes, sample sizes, and alternative approaches to the ANOVA. As he notes, the results of his analysis "have implications for many other fields of study." This well-conceived and carefully reasoned treatise represents a significant contribution to our understanding of obstacles that may prevent or deter the obtaining of valid scientific knowledge, with respect to the heredity-environment controversy and more generally.

It seems to this commentator that Wahlsten has provided the conditions for a more definitive solution to the problem. This can be accomplished by designing an appropriate computer simulation which would provide answers to a number of critical questions that still plague the field.

One could begin perhaps with two general conditions to simulate the H·E "true" state of affairs: 1. The Null Case, one in which an ideal additive population is simulated (i.e., absence of an H·E interaction effect), and 2. The Non-Null Case, in which an ideal nonadditive population is simulated (i.e., presence of an H·E interaction effect that is both statistically and biologically meaningful). Other sources of variation that could be studied systematically include the specific type of interaction deriving from the simulated population; the type of statistical approach to data analysis (e.g., ANOVA; the Neyman or Traxler, 1976, approach; the likelihood ratio approach); the simulated effect size itself (small, medium, large).

Once these populations are constructed, one could then draw randomly from them (with replacement) in the sample size range usually used in research studies of H·E effects. It would then be a relatively straightforward matter to calculate, under each of the simulated conditions (or assumed "true" states of knowledge), at least the following: alpha (Type I) error, beta (Type II) error, power, effect size, strength of the relationship, and minimal sample size requirements for detecting correctly the population results.

It should be realized that computer simulations appropriately designed and executed, can be used to answer a host of important questions that cannot be answered on the basis of a small number of isolated studies. Thus, for example, my colleagues and I have used computer simulation methodology to understand why so many inconsistencies and contradictory findings exist among hemispheric-asymmetry studies involving left-handers. In comparing samples randomly drawn (with replacement) from a large parent population of sinistrals, we assessed the role of alpha and beta errors in inconsistent results in the field. We could distinguish sample and population results that were statistically meaningful from those which were both statistically and clinically relevant, and we realized in a much broader context the need for a more positive attitude toward the design, execution, and publication of replication studies (Soper et al. 1988). Clearly, these are all issues for H·E research as well.

Given that the analysis of covariance (ANCOVA) is a close relative of ANOVA, the author should consult the scholarly contribution of Adams et al. (1985), who used computer simulation methodology to examine some of the assumptions underlying ANCOVA (analysis of covariance), as they relate to research in the field of neuropsychology.

In summary, then, the Soper et al. (1988) and the Adams et al. (1985) investigations can be explored for general information concerning the design of computer simulation studies. The author could, if desired, tailor his own research design to fit the specific requirements for further investigations of the H·E interaction issue.

How important is detecting interaction?

James F. Crow

Genetics Department, University of Wisconsin, Madison, WI 53706
Electronic mail: *wrengels@wiscmacc.bitnet*

Wahlsten's point about the difficulty of detecting significant interaction of heredity and environment is well taken, but I disagree that it is meaningless to measure heritability unless the additivity assumption is met. A simple genetic example, similar to the gravitational model that he discussed, illustrates the point. Suppose that the true relation between genetic and environmental factors is multiplicative, but we don't know this and follow the usual procedures, analyzing the data as if the factors were additive. How great is the error of estimation of heritability and how likely are we to detect the interaction?

Consider a trait determined by four equal, additive, normally distributed genetic factors and four environmental ones. The trait value is the product of the two sums. To model parent-offspring correlation in random environments, two of the four genetic elements are shared by the two relatives. A simulation generated a correlation of 0.242, leading to a heritability estimate of 0.484, close to the 0.5 expected from the model of equal genetic and environmental influences (Table 1, line 1). The error is minor despite a rather wide range of values; the coefficient of variation, 0.36, is larger than that for such quantitative measures as size, blood pressure, and IQ. The numbers in line 2 are based on the perhaps more realistic assumption that factors are multiplicative within as well as between the two components. The heritability is slightly farther from the expected value, as expected, but still very close.

Table 1 (Crow). *The effect of curvilinearity on estimates of heritability from parent-offspring correlation*

Line	Transformation	Mean/ Maximum	Coef. Var.	Est. h^2
1	X_1X_2		0.36	0.484
2	e^x		0.36	0.474
3	e^x		0.53	0.459
4	$\log_e X$		0.40	0.491
5	tanhX	0.591	0.37	0.476
6	tanhX	0.755	0.31	0.438
7	$2X - .3X^2$	0.499	0.30	0.495
8	$2X - X^2$	0.815	0.30	0.317
9	$2X - X^2$	0.876	0.20	0.123

What would happen if a breeder naively used the observed parent-offspring correlation measures to predict the result of selection? In line 2 the mean is 7.856, with a standard deviation of 2.862. Suppose the upper 50 percent of his population are permitted to reproduce. If the breeder erroneously assumes normality and additivity, this corresponds to a selection differential of 0.80; that is, the mean of the selected group is 0.8 standard deviations above the population mean. The heritability has been estimated as 0.474, so the progeny will be $0.8 \times 2.862 \times 0.474 = 1.085$ units above the mean, or 8.941.

Suppose the breeder is more sophisticated and applies a log transformation to these data. The transformed mean and standard deviation are 2.000 and 0.354, and the calculated heritability is 0.497 (very close to the theoretical 0.5). The expected increase is $0.8 \times 0.354 \times 0.497 = 0.1407$, so the progeny mean (m) and standard deviation (σ) are 2.1407 and 0.354. Our sophisticated breeder of course transforms back into the original units. These have a log-normal distribution and the transformed mean is $exp(m + \sigma^2/2)$, or 9.055 (Moran 1968, p. 317). The error of predicting the next generation by the naive procedure, 8.941 versus 9.055, is only about 1% – not worth worrying about.

The interaction induced by multiplicative behavior in this case is very small, and could never by detected by any reasonably sized experiment, even if there were not the uncontrolled errors and covariances that beset such studies. But the positive side is that the error introduced by ignoring interaction is trivial.

1. Transformations to remove metrical bias. It is convenient to classify interactions into two categories. The first kind results from what R. A. Fisher called metrical bias and usually preserves rank order. This can often be removed by a simple transformation. The second kind, perhaps more interesting, is too complex to be removed by any simple or obvious transformation. The multiplicative example is the first kind.

Wahlsten fears that a transformation may conceal interesting relationships, and it may. Yet, it seems to me that the discovery of a linearizing transformation usually tells more about the underlying mechanisms than would the finding of an interaction component of the variance alone. That a log transformation is linearizing does not, of course, prove that genes and environmental factors multiply, but that's the way to bet. It provides a hypothesis to be tested further.

Table 1 gives some additional examples. In each case, I generated measurements by adding 8 normally distributed, random variables and then transformed those measurements into what would correspond to observed values. The transformation is, of course, the inverse of what would be used to linearize the observations.

Line 3 shows the effect of an enhanced range; the coefficient of variability is increased to 53 percent, and the estimated

heritability is now about 92% of its expected value. Line 4 considers data that interact in the opposite way; the curvature is downward rather than upward. Lines 5 and 6 model a situation in which the distribution saturates at the upper end. Again the error is not great, but becomes greater as the mean approaches the maximum, as expected.

Plant and animal breeders, as well as human quantitative geneticists, often neglect interactions because there is no practical way to measure and deal with them. An appropriate transformation can improve the accuracy of prediction, but often the game isn't worth the candle. If the interaction is due to metrical bias, other errors are likely to dwarf those due to nonadditivity.

On the other hand, there may not be any simple transformation that linearizes the process. This kind of interaction – *real* interaction – calls for a different approach. Unfortunately, simply showing that there is an interaction component to the variance doesn't suggest how to go about understanding it. The approach depends on the subject matter. An interaction component in the analysis of a series of corn hybrids might suggest testing the influence of specific variables such as amount of rain, day length, and temperature. This might reveal that some strains are best suited to a long growing season with heavy rainfall, for example, and would permit allocating strains to specific environments. Behavioral studies designed to identify interacting factors have to be based on some hypothesis and an experimental design appropriate to it.

One kind of relationship that can lead to greatly reduced heritability occurs when there is an intermediate maximum value for the trait; beyond a certain point, genetic and environmental factors that previously increased the trait now decrease it. Again as expected, the nearer the mean is to the maximum, the lower the heritability. The last three examples in Table 1 are of this type. They differ from the first six in that they cannot be linearized by a transformation; no single-value inverse function exists. This kind of relationship could lead to a serious underestimation of the degree of genetic influence; but a (narrow sense) heritability analysis would correctly predict the slow progress of breeding experiments. In this case, improving the environment wouldn't help much either. The interaction must be taken into account, if it can be understood.

2. What use is a heritability estimate? I disagree with Wahlsten's view that the only reason to know heritability is to predict the results of selection. One can be interested in the heritability of IQ without advocating a breeding program. A broad sense heritability of 0.2, based on identical genotypes in independent environments, tells us that 80 percent of the variance is environmental. This says that changes of existing variables within the existing range can have a substantial effect on the trait. On the other hand if the heritability is high, environmental manipulations will have little effect unless they extend beyond the existing range, or bring in new factors.

A strength of a heritability estimate is that, remarkably, it tells us how much influence existing environmental factors have, even when we have no idea what the environmental factors are. This is also its weakness, for one has to look elsewhere to discover *which* environmental factors are important. Unfortunately, environments don't follow the simple, mechanistic rules of Mendelism. It is a matter of knowledge and ingenuity on the part of the investigator to pick out likely candidates and test them.

I have the same view of interaction. Just detecting genotype-environment interaction tells us little. One has to have specific testable hypotheses. To increase the probability of finding interactions, increasing the size of the experiment is often too expensive. It may be more profitable to enhance the genetic variance by, for example, using diverse inbred lines (which, incidentally, reduces the error variance), and to manipulate the environment by extending the normal range or introducing novel factors.

3. Why have laboratory experimenters and those involved in uncontrolled breeding or population experiments treated interaction so differently? Wahlsten finds this surprising; I don't. Drosophila and mouse mutations with major effects often interact in complex ways. The study of embryology using genetic methodology emphasizes interactions, and the interactions of mutations have often been the key to a deeper understanding. An excellent example is the molecular and mechanistic understanding of the nervous system that has come from just such studies. The whole pattern of embryological development thus revealed is a tangled network of interactions. The interaction of specific mutations with specific environmental agents – viruses or hormones – has likewise yielded fruitful insights.

If interactions are so important, why have seemingly naive additivity assumptions been so successful in breeding experiments? Why has selection worked as well as it has? Why have its results been predictable from simple theory? Part of the answer, of course, is that the sizable uncontrolled errors make it hard to detect departures from predictions. But I think there is also a deeper reason.

Performance traits in livestock, and very likely the genetic components of such things as human intelligence, are determined mainly by the combined effects of many allelic differences with individually small effects. For the same reason that linear approximations work in physical sciences (Hooke's Law, the linear terms in a Taylor expansion), these small effects are approximately additive. Tiny increments of *anything* are additive. This includes environmental as well as genetic factors. A similar conclusion comes from the findings of Keightley (1989) on enzyme kinetics. In his words, "Data on enzyme activity variation from natural and artificial populations suggest that such variation generates little nonadditive variance despite the highly interactive nature of the underlying biochemical system."

To go further afield: Evolutionists have long agreed that the steady improvement of adaptation depends mainly on multiple genes with small effects. If interactions were rampant, evolution (at least in sexual species) would be impossible. The results of selection would be chaotic. A certain amount of additivity is a prerequisite for evolution.

So I don't find the difference in emphasis surprising. If the object is to make predictions and determine the relative importance of genetic and environmental influences in the existing population, there is sufficient additivity to render correlation and variance analysis appropriate. If the idea is to get at deeper mechanisms, then one needs to identify specific genes and environmental factors with large effects, not part of the normal population; they are quite likely to interact, and finding strong interactions argues that they are functionally related.

Eddington, quoted by Fisher (1930, p. viii), said: "We need scarcely add that the contemplation in natural science of a wider domain than the actual leads to a far better understanding of the actual." That is what experimenters do. They introduce mutant genes that are far outside the range of anything that would survive in nature and they consider environmental extremes far outside the ordinary, or even of a totally different kind. This is the way to find causes and interactions among causes.

Estimating heritabilities in quantitative behavior genetics: A station passed

Wim E. Crusio

Institut für Humangenetik und Anthropologie, Universität Heidelberg, 6900 Heidelberg, Federal Republic of Germany
Electronic mail: *j31@dhdurz2.bitnet*

Wahlsten's target article boils down to an eloquent attack on quantitative behavior genetics and in particular on the practice of partitioning the variance among mutually exclusive causes (heredity and environment). He focuses on genotype-environ-

ment interaction (G×E), but he might just as well have chosen genotype-environment covariation as the target for his wrath. When written in full, quantitative-genetics' basic causal model is

$$\mathrm{Var(Y)} = \mathrm{Var(H)} + \mathrm{Var(E)} + 2\,\mathrm{Cov(H,E)} + \mathrm{Var(G{\times}E)}$$

(assuming that G×E is uncorrelated with either H or E; see Plomin et al. 1980). Partitioning the variance into components due to hereditary and environmental causes is thus valid only if *both* interaction and covariation of genotype and environment are negligible. This condition can be approached only in tightly controlled animal experiments, where subjects with different genotypes may be provided with as uniform an environment as technically feasible. In such a situation, only microenvironmental variation could induce either interaction or covariation between genotype and environment. Here, G×E may be adequately dealt with by applying appropriate transformations. Wahlsten quite justifiably remarks that "the interpretation . . . is altogether different . . . if we forget about the transformation." Forgetting important features of an experiments' design or analysis is obviously a bad research strategy. In any case, if either interaction or covariation between genotype and environment (or, for that matter, interactions between genotype and treatment, environment and sex, etc.) are present in a particular experiment, estimates of heritabilities are uninterpretable.

Still, as noted by Wahlsten, even when heritability estimates are valid, they can only be used to predict the effects of possible selection pressures. As such, these estimates have only a very limited value for researchers investigating animal behavior and are without purpose in human-behavior research. Studies whose only goal it is to estimate the heritability of a psychophene have been useful in the past when many ethologists and psychologists had to be convinced that heredity can play an important causal role in interindividual differences in behavior. It is by now clear that this approach is basically sterile and that these efforts should be abandoned.

The above applies, also of course, to most behavior-genetic analyses in which variance is partitioned. Although the knowledge that a significant portion of phenotypic variation is attributable to a certain source might be of interest (e.g., Plomin & Daniels 1987), information about the absolute or relative size of this portion in a certain experimental situation is less important (because it will vary with the specific conditions of that particular situation).

All this should not be interpreted as a plea to abandon all quantitative-genetic procedures relying on the partitioning of variance. As explained recently (Crusio in press; Crusio et al. 1989), combining these methods with multivariate techniques might allow the analysis of causal relationships between, for example, neural and behavioral phenotypes. Estimating heritabilities as a goal in itself is clearly a thing of the past. We should now occupy ourselves with more important and more rewarding problems.

Monotone interactions: It's even simpler than that

Robyn M. Dawes

Department of Social & Decision Sciences, Carnegie Mellon University, Pittsburgh PA 15213-3890

Electronic mail: *bitnet: rd1b@andrew.cmu.edu*

I agree, completely. My only concern is that the complexity of the presentation – involving simulations and analogues – will obscure the simplicity of the basic point. With the editor's indulgence, I would like to present a simpler explanation, which I have used with my students for many years.

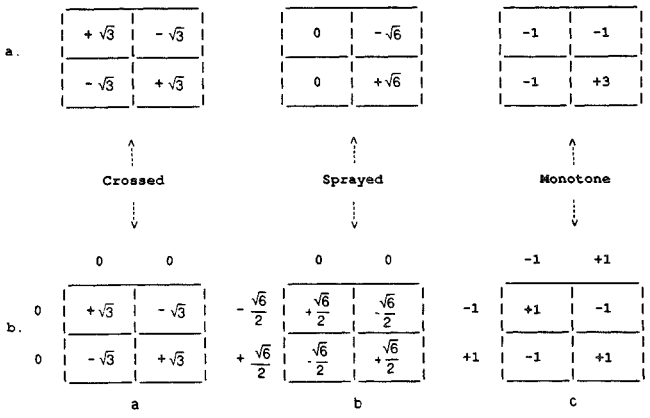


Figure 1 (Dawes). The three 2 × 2 interactions.

Figure 1 presents three interactive patterns of group means in a 2 × 2 design in which the sum of squares between groups is constant (= 12). The first pattern [1(a)] is termed *crossed* (corresponding to Wahlsten's "Case 2" in his figure 2(a)); the second [1(b)] is termed *sprayed* (corresponding to Wahlsten's "Case 1"); the third is termed *monotone* (corresponding to Wahlsten's Y = H·E example). Since there are four means, any interactive patterns in the 2 × 2 design can be conceptualized as a linear combination of these three basic patterns. (Indeed, with an arbitrary zero point – here set equal to the overall mean – any pattern at all can be, including a "no interaction" pattern.)

The bottom part of Figure 1 breaks down the deviations (from the overall mean of 0) into row effects (presented on the side), column effects (presented at the top) and interaction effects (presented in the cells) according to the standard ANOVA model. The important result is that the sum of squares for interaction for a sprayed interaction is one half that of the corresponding crossed one, and for a monotone interaction one-third. (That reflects the "overall effect" of a sprayed interaction's being split equally into row and interaction effect and that of a monotone interaction's being equally distributed across row, column, and interaction effects.) It follows directly from the definition of the standard F-test that three times the sample size is required to obtain a specified significance level for a monotone interaction as for a crossed one of the same magnitude – where magnitude is defined in terms of the sum of squares between means. Twice the sample size is required for a sprayed interaction. If magnitude is defined in terms of the sum of absolute deviations, rather than the sum of squared deviations, the comparable proportionality figures are 9 and 4.

It is difficult for investigators to increase sample sizes by such radical amounts when they realize that the "interactions" they have hypothesized (usually verbally) are not crossed. Many (at least several of my students and colleagues) then have the creative idea of analyzing their data through multiple regression techniques using dummy coefficients "fitted" to the type of interaction anticipated. Thus the row and column factors for a 2 × 2 crossed interaction are coded (+1, −1) (so that the upper left and lower right cells are weighted +1, the others −1); the row and column factors for a 2 × 2 monotone interaction are coded (0, 1) (so that one product is 1, the other three 0), and those for a sprayed interaction are coded (0, 1) and (+1, −1). The (again verbal) rationale is that while ANOVA tests for interactions *of any sort* through the omnibus F-test, the more specific coefficients provide tests of particular types of interaction. The analogy is often drawn to testing specific trends rather than relying on omnibus tests – e.g., in contingency tables or one-way ANOVA designs.

Unfortunately, this procedure does not work – at least not for the 2 × 2 design.[1] The problem is, as Cohen and Cohen (1975, pg. 295) point out: "Only when *u* [a main effect] and *v* [the other main effect] have been linearly partialled from *uv* does it, in

general, become the interaction IV we seek." Let me illustrate.

Consider a monotone pattern of cell means (3, 3, 3, 9) that correlate perfectly with the coefficients (0, 0, 0, 1) of the interaction term obtained when the main effects are coded (0, 1). The multiple correlation from both main effect coefficients and interaction coefficients is 1.00, but the main effect coefficients each correlate .577 with the cell means, and since they are independent, they jointly predict the cell means with an R^2 = .66 (= $.577^2 + .577^2$). The residual increase in R^2 from the interaction coefficients is thus .33 (= $.577^2$). In contrast, consider the coefficients (+1, −1, −1, +1) of the interaction term obtained when the main effects are coded (−1, +1). Again, the multiple correlation is +1.00, and since the main effect correlations are unaltered, the residual increase in R^2 for the interaction coefficients is again (.33 (= $.57^2$). The only difference is that with this (−1, +1) coding the interaction coefficients are uncorrelated with the coefficients of the main effects. The residual increase in R^2 is identical.

Conversely, consider a crossed pattern of cell means (3, 0, 0, 3) that correlates perfectly with the interaction coefficients (+1, −1, −1, +1) obtained from the (−1, +1) coding of main effects. How can the "residual" correlation of the interaction coefficients (0, 0, 0, 1) – obtained from (0, 1) coding of the main effects – also have a residual value of +1.00? The answer is found by noting that the (3, 0, 0, 3) pattern is perfectly predicted by weighting each main effect −3, the interaction effect +6, and adding +3 to the result; for example, $3 = (-3) \times 0 + (-3) \times 0 + (+6) \times 0 + 3$; $0 = (-3) \times 0 + (-3) \times 1 + (+6) \times 0 + 3$, and so on. Thus, the cell means are perfectly predicted from a linear combination of the main and interaction effects. Since the correlation of both main effects with (3, 0, 0, 3) is zero, however, the residual effect of the (0, 0, 0, 1) interaction coefficients accounts totally for this perfect correlation. What has happened is that the main effects have become *suppressor* variables, since their coefficients are positively correlated with the coefficients of the (0, 0, 0, 1) interaction term but uncorrelated with the cell means.

My illustration involves perfect prediction in the simplest 2 × 2 design. The general principle applies *mutatis mutandis* to imperfect prediction in more differentiated designs. The general linear model is a *general* linear model.

Finally, students and colleagues sometime confuse interaction with confounding – as would occur in Wahlsten's context if intelligent organisms selectively migrated to "intelligent environments." That's a whole different problem, a complex one involving the relationship between "balanced" and "representative" designs – and futile (Darlington 1968) attempts to unconfound intrinsically confounded variables in contexts not involving experimental manipulation.

We are stuck with the profound 66-year-old conclusion of Fisher and Mackenzie (1923) that interaction is whatever is left over after main effects are removed. Everything Wahlsten pointed out follows from that observation.

NOTE

1. Over my objections, one proposal using such coefficients has been published in a select journal in the psychological literature. Rather than cite it to tear it apart, however, I warn the reader that such propositions occur – in the most reputable settings.

Effects of correlation on interactions in the analysis of variance

Victor H. Denenberg

Biobehavioral Sciences Graduate Degree Program, University of Connecticut, Storrs, CT 06269-4154
Electronic mail: *biosadm3@uconnvm.bitnet*

Several years ago two graduate students showed me a set of data represented by two curves. Curve A increased linearly over the five observation periods. Curve B was flat over the first three trials and was numerically lower than A; it then increased very sharply and had higher values than A for the last two trials. It was "obvious" to me and to the students that there had to be an interaction of treatments (A or B) and trials since the two curves crossed. But the students had found the interaction to be numerically small and far from significant. They were puzzled and asked me for an explanation. I didn't have one at the time, and put the problem aside to think about at some future date. Wahlsten's fine paper, with his demonstration of the loss of power for tests of significance of interaction in the analysis of variance (ANOVA), was the necessary stimulus to get me to return to that problem.

Table 1 (Denenberg). *Two group repeated measurement design*

Subject	C	E	D = E − C
S_1	4	7	3
S_2	6	7	1
S_3	8	7	−1
S_4	10	12	2
S_5	12	14	2
s^2	10.00	11.30	2.30
s	3.16	3.36	

$r_{ce} = .8937$

$t = \text{Mean D}/SE_d = \text{Mean D}/(s_d^2/N)^{1/2}$
$= 1.40/(2.30/5)^{1/2} = 2.06$

Source	df	MS
Trt	1	4.90
S	4	20.15
TS	4	1.15

$F = MS_t/MS_{ts} = 4.90/1.15 = 4.26$
$t^2 = F = (2.06)^2 = 4.24$
$s_d^2 = s_c^2 + s_e^2 - 2r_{ce}s_cs_e$
$= 10.00 + 11.30 - (2)(.8937)(3.16)(3.36) = 2.30$
$= 2nMS_{ts} = 2.30$

I approached the interaction question from a vantage point different from Wahlsten's, and without any initial concern about power. However, my solution led to the same general conclusion, although with an important restriction. My general thesis is that the correlation between the treatment conditions (in my example Curves A and B) significantly affects the numerical value of the interaction term, and this may result in the loss of statistical power. Let me now develop that argument.

To begin, consider the data in Table 1. Five subjects were tested under a control and an experimental condition. The first three columns contain the subject identification and their data. The last column is the difference between the C and E scores and will be used to calculate the correlated *t*-test. The column variances and the standard deviations for the C and E columns are given below the table. Beneath that is given the correlation between C and E. These data will be used in the calculations that follow.

The test of the mean difference between C and E can be done via a correlated *t*-test or the analysis of variance. Both approaches will be used here. The easiest way to do the correlated *t*-test is to use the difference score approach as discussed in any introductory statistics text. The calculations using the difference scores are shown in Table 1, and *t* is found to be 2.06. Using ANOVA, the *F*-test for the treatment effect is 4.26. With 1 *df* in the numerator $t^2 = F$. In this problem $t^2 = 4.24$ which is within rounding error of 4.26.

The next point to make is that the variance of the difference scores, s_d^2, can be obtained from the data in the C and E columns and is a joint function of the variances of these two columns and their correlation. The formula is

$$s_d^2 = s_c^2 + s_e^2 - sr_{ce}\, s_c\, s_e \quad (1)$$

The correlation between the C and E columns is .8937. When Formula (1) is applied to the data in Table 1, $s_d^2 = 2.30$, which is identical to s_d^2 when directly calculated.

One more point is needed to complete my argument. Note that the error term for the ANOVA, MS_{ts}, is an interaction term and has the value 1.15. Also note that s_d^2, used in the denominator of the *t*-test, has the value 2.30, twice that of MS_{ts}. The general relationship between any interaction, MS_{int}, and s_d^2, is:

$$2nMS_{int} = s_d^2, \quad (2)$$

where n = number of observations per cell, and the calculations are carried out in the usual manner of using totals rather than means. In this example, n = 1. The last line in Table 1 shows this relationship.

We can combine Formulas (1) and (2) to define MS_{ts} as follows:

$$MS_{ts} = [s_c^2 + s_e^2 - 2r_{ce}\, s_c\, s_e]/2n \quad (3)$$

We now see that the interaction term is directly affected by the degree and direction of correlation between the two treatments (C and E). This is the major point of my thesis. However, before drawing any general conclusions, it is necessary to show that the relationships described above are not limited to the correlated *t*-test but also apply to factorial experiments.

The single-factor repeated measurement design in Table 1 can easily be generalized to a factorial experiment. Table 2 shows a 2 × 4 factorial experiment with 4 observations per cell. Only the cell totals are given. From these one can calculate the mean squares for A, B, and the AB interaction. (The error mean square is irrelevant and is not shown.) In addition to the usual information, the final column of that table shows the difference between the A_2 and A_1 treatments, in parallel with Table 1. Beneath the data table are listed column variances and the standard deviations for A_1 and A_2. The correlation between A_1 and A_2 is .3843. MS_{ab} is found to be 155.88 from the ANOVA. The variance of the difference between A_1 and A_2, s_d^2, is 1247, whether calculated from the difference column in Table 2 or via Formula (1). This is equal to $2nMS_{ab}$.

Table 2 (Denenberg). 2 × 4 *Factorial design.*

	A_1	A_2	$A_2 - A_1$
B_1	23	45	22
B_2	80	97	17
B_3	110	58	−52
B_4	90	65	−25
s^2	1392.25	488.92	1247.00
s	37.31	22.11	

n = 4 per cell

$r_{a1.a2}$ = .3843

Source	df	MS
A	1	45.12
B	2	314.42
AB	2	155.88
Error	24	

$s_d^2 = 1247.00$
$= s_{a1}^2 + s_{a2}^2 - 2r_{a1.a2}\, s_{a1}\, s_{a2} = 1247.13$
$= 2nMS_{ab} = 1247.04$

Table 3 (Denenberg). 2 × 3 *Factorial design.*

	A_1	A_2	$A_2 - A_1$
B_1	24	31	7
B_2	9	31	22
B_3	49	25	−24
s^2	408.33	12.00	550.33
s	20.21	3.46	

n = 3 per cell

$r_{a1.a2}$ = −.9286

Source	df	MS
A	1	48.39
B	2	1.39
AB	2	91.72
Error	12	

$s_d^2 = 550.33$
$= s_{a1}^2 + s_{a2}^2 - 2r_{a1.a2}\, s_{a1}\, s_{a2} = 550.20$
$= 2nMS_{ab} = 550.32$

A final example is given in Table 3 for a 2 × 3 factorial with 3 observations per cell. Here, however, the correlation between A_1 and A_2 is −.9286, and this results in a very large interaction term.

I can now answer my graduate students' question. The reason for the failure to find the "obvious" interaction was that the two curves had a high positive correlation, and this correlation acted to reduce the variance associated with the interaction mean square, as can be seen from inspection of Formula (3). This is a perfect example of Wahlsten's thesis that the test for interaction has less power than the test for main effects.

Since MS_{int} and s_d^2 are functionally related, this means that Formula (3) can be used to help gain insight into Wahlsten's arguments. The crux of the formula resides in the correlation term, r_{ce}. When the correlation is positive, the value of MS_{int} is reduced; when negative, it is increased. This leads to the general conclusion that *an interaction mean square will be numerically reduced if the correlation between the two data sets is positive; it will be increased if the correlation is negative, and it will not be affected if the correlation is zero.* Reducing the value of the interaction term (relative to the zero-correlation condition) is equivalent to reducing the power of the interaction test. Thus we arrive at Wahlsten's position. However, according to my development, his conclusion holds only for the condition where the correlation between two data sets (or the average intercorrelation when there are more than two data sets) is positive. In contrast, a negative correlation would act to increase the power of the interaction test.

Finally, I enthusiastically applaud Wahlsten's emphasis on the use of interaction as a sign to the investigator to think carefully about the meaning of underlying processes. Statistics, like other tools, is to be used in the service of science, not as an end in itself.

Don't kill the ANOVA messenger for bearing bad interaction news

Douglas K. Detterman

Department of Psychology, Case Western Reserve University, Cleveland, OH 44106

Wahlsten makes two serious errors in his analysis of the use of interaction in behavior genetics. First, he equates science with statistics. Second, he fails to take into account previous attempts to discover H x E interactions. More seriously, he compounds these two errors by implying that behavioral genetic analyses, as currently carried out, are seriously flawed.

The first problem with Wahlsten's analysis is that he equates the statistical methods used with the theory. This seems to be a frequent error of the statistically sophisticated and it is particularly prevalent among students who have just completed a series of statistics courses. Statistics are merely a tool and have nothing whatsoever to do with the correctness of a theory.

Behavior genetics makes a clear statement about genetic (H) and environmental (E) influences related to phenotypic (P) variance:

$$P = H + E + HE + \text{error}$$

Analysis of variance, on the other hand, makes a clear statement about the score model when there are two variables in the model:

$$X = A + B + AB + \text{error}$$

Because there is a superficial correspondence between the two, it is possible to use analysis of variance to test the behavior genetic model. But the statistical model is different from the model proposed by the theory. The validity of the behavior genetic model is completely unaffected by the power of analysis of variance. Analysis of variance is only a tool and, as such, is no more relevant to the substance of a theory of behavior genetics than are arguments over the refining powers of optical and atomic microscopes to theories of cell development. Though one method may give better kinds of data to support or reject the theory, the method itself has no bearing on the essence of the theory.

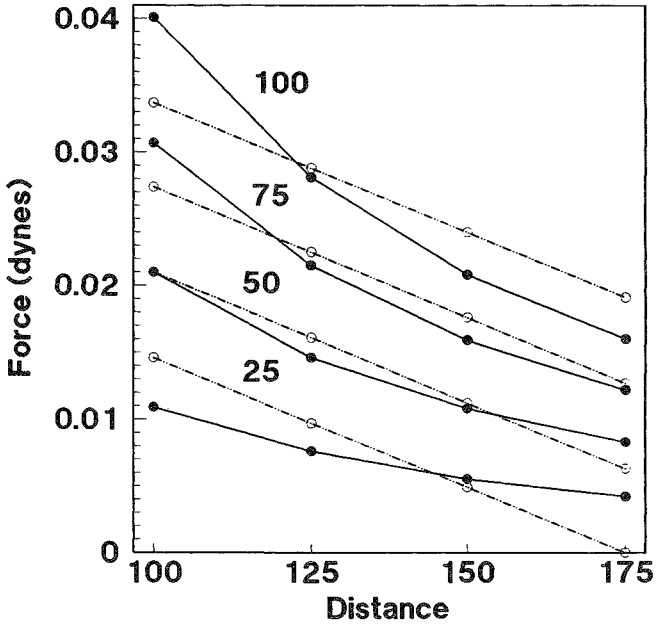


Figure 1. (Detterman). Predicted values of Wahlsten's sample gravity data without an interaction term. Dashed lines are predicted values, solid lines are actual values for each mass of m_2.

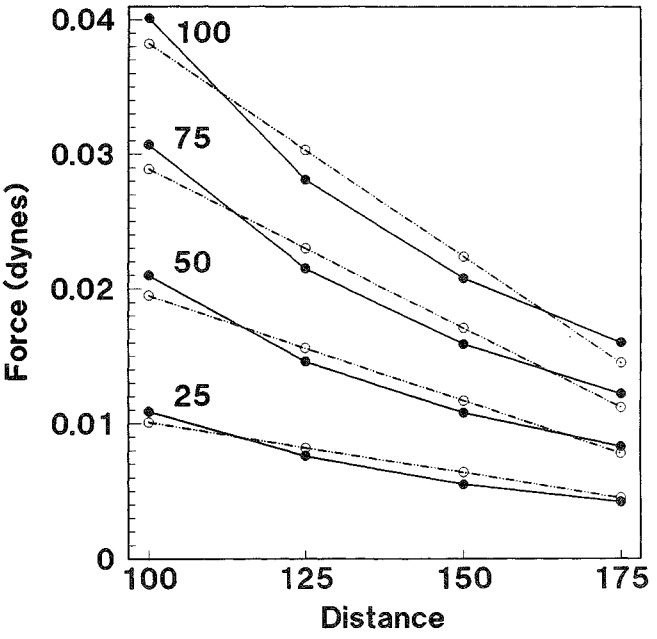


Figure 2. (Detterman). Predicted values of Wahlsten's sample gravity with an interaction term. Dashed lines are predicted values, solid lines are actual values for each mass of m_2.

This point is easy to see in a reconsideration of data which Wahlsten considers to be an instructive example: The gravitation data is presented to show the shortcomings of analysis of variance methodology, but this example is a better demonstration of the shortcomings of the hypothetical scientist who collected the data than of the methodology used. Wahlsten generated data from the equation giving the mutual force of attraction between two bodies. He then analyzed those data with analysis of variance and, because the interaction between mass and distance was not statistically significant, concluded that the analysis of variance would fail to capture the law of gravitation.

To determine exactly what went wrong, I used multiple regression to predict the group means of force from mass and distance first and then from mass, distance, and the mass X distance interaction. Figure 1 shows the predicted scores from a regression equation using only mass and distance. The squared multiple correlation is .89. Figure 2 shows the predicted scores from a regression equation using mass, distance, and the interaction of mass and distance. The squared multiple correlation is .96. The addition of the interaction terms accounts for about 7% of the total variance. Graphically, it is easy to see that the fit in Figure 2 is only slightly better than the fit in Figure 1 *for these data*. In addition, both fits are better than you would even get from most theories in the behavioral sciences.

Science requires parsimony – an explanation is most powerful when it is simplest. If the interaction effect of mass X distance in the above analysis were not statistically significant, parsimony would require us to accept the simple additive model of mass and distance. This simple model provides a highly adequate fit of the data at hand. Analysis of variance is providing an appropriate conclusion for the data set used.

However, I do not think that any reasonable scientist could look at Figure 2 and not long for more data on the question. The obvious problem, acknowledged by Wahlsten, is that the range of values is restricted. Had a wider range of data been used it is obvious the conclusion would have been quite different. Statistical techniques, like computers, follow the GIGO principle (garbage in, garbage out).

From Wahlsten's analysis of these data, he would have us conclude that we should all be thankful that Newton never invented analysis of variance. But I conclude from my reanalysis that we are fortunate that Newton collected data more sensibly

than Wahlsten constructed his examples. It is certainly not possible to conclude from Wahlsten's analysis of variance that the law of gravitation is incorrect. Neither can it be concluded that the analysis of variance is flawed. The only appropriate conclusion would seem to be that the data collected must be appropriate for testing a specific theory with a given method. In that respect, I fully agree with Wahlsten's conclusion that larger numbers of subjects should be used in such studies. But I already believed that before reading Wahlsten's analysis (Detterman 1989).

The second problem with Wahlsten's argument is that he gives little attention to the previous efforts to find H x E interactions. Plomin (1986), on the other hand, devotes an entire chapter to the topic, considering both human and animal studies. And Plomin is well aware of the power issues involved in the detection of interactions (see p. 106). Plomin's conclusion, based on this review, was that H x E interactions are hard to find and when they exist they account for a small portion of the variance. This is remarkably similar to the conclusions drawn by Cronbach and Snow (1975) regarding aptitude by treatment interactions. Evidently, interactions between abilities and environment do not account for a large portion of the variance in human behavior.

Wahlsten *must* agree that H x E interactions do not account for a large portion of the total phenotypic variance. If H x E interactions accounted for larger portions of the variance then even statistical techniques with lower power would be able to detect them. If the H × E effects are very small, then the question becomes one similar to the gravitation example above: Is it more parsimonious to exclude or include the interaction effect? Currently, researchers in behavior genetics find no compelling reason to include interactions in their models. But like any group of scientists they can easily be convinced by overpowering data. The real problem with H x E interactions is not with the analysis of variance but with the absence of persuasive data indicating the interactions are important to a behavior genetic model. I would suggest that if Wahlsten believes that H x E interactions are important and he knows why they have not been found, that he carry out research demonstrating some important H x E interactions. Such findings would make an extremely important contribution.

Finally, there is an implication running throughout Wahlsten's target article that a variance accounting approach should be abandoned because, as nearly as I can tell, the method involves the computation of heritability. Wahlsten seems to believe that heritabilities are somehow evil, noting that "the only practical application of a heritability coefficient is to predict the results of a program of selective breeding." This is the emotive equivalent of saying that the only practical reason for studying cholera is to develop germ warfare weapons. Heritabilities are nothing more than a way of representing portions of variance according to a particular model. In my opinion, a variance accounting approach also has many desirable characteristics which recommend it over a hypothesis testing approach. In short, the arguments presented by Wahlsten are not convincing either against the use of analysis of variance or against the currently prevalent theory of behavior genetics.

ACKNOWLEDGMENT

Parts of this work were supported by Grants No. HD07176 and HD15516 from the National Institute of Child Health and Human Development, Office of Mental Retardation, the Air Force Office of Scientific Research and the Brooks Air Force Base Human Resources Laboratory, Project Lamp.

Interaction and dependence prevent estimation

R. M. Dudley

Mathematics Department, Room 2-245, Massachusetts Institute of Technology, Cambridge, MA 02139

In human behavior genetics, IQ has been studied perhaps more than all other traits combined (Plomin 1983). The data are usually summarized by kinship correlations (Bouchard & McGue 1981), with special emphasis on twins and adoptive kinships. There are interaction problems even thornier than those Wahlsten rightly points out. Very few analyses of variance as such for IQ have been published; they cannot be done from correlations alone.

Interaction can be defined in theory without reference to analysis of variance. Let S be a function of genome and environment, here an IQ test score. In any given population under study, genomes and environments each have probability distributions. Let C be the average value of S if an individual's genome and environment were chosen independently at random. For a given genome, let the average value of S over environments be C + G, the *genotypic value*. For a given environment, let the average values of S over genomes be C + E, the *environmental value*. The additive model holds if $S \equiv C + G + E$. In general, $S = C + G + E + I$, where I is the *interaction*. For a further analysis, the interaction can be expanded as

$$I = G_1E_1 + G_2E_2 + \ldots,$$

where G_i are genetic and E_i are environmental variables (Freeman 1973), all with 0 averages. G_i need not be functions of G, nor E_i of E.

"Heritability" is defined (here) as the ratio of variances $h^2 = \text{var}(G)/\text{var}(S)$. Its dependence on the particular distributions of genomes and environments under study is too often neglected. The IQ data base spans cognitive environments varying with time (across the advent of television), geography, etc.

For humans (and for other species in their normal habitat outside of designed studies), environments and genomes are dependent (family members share genes and environments). Then the variance of S is a sum of variances and covariances,

$$\text{var}(S) = \text{var}(G) + \text{var}(E) + \text{var}(I) + 2(\text{cov}(G,E) + \text{cov}(G,I) + \text{cov}(E,I)).$$

It is impracticable to estimate both var(I) and cov(G,E) if they are not zero (Layzer 1974). The terms cov(G,I) + cov(E,I) raise further obstacles. They can be removed (Jacquard 1983) by changing definitions, which does not solve the original problem.

Even under an additive model S = C + G + E, and even if E is uncorrelated for any two individuals, possibly living together, the estimation of heritability from kinship correlations is complicated by genetic phenomena: dominance, epistasis (nonadditive interactions within genetics), assortative mating (Carter 1977) and lack of equilibrium (Ewens 1979, p. 287). Each G_i in the interaction is subject to these complications. Assortative mating no doubt also has environmental aspects.

The fit of the additive model to data can be deceptive because interaction effects on kinship correlations can be confounded with purely genetic effects (Le Roy 1960a, p. 122; 1960b; Plomin et al. 1977). Specifically, one-egg twins raised together will have interaction terms more alike than those of two-egg twins, a complication in the "twin method." If the additive model "explains" more than 100% of the variance (Jencks et al. 1972, p. 266) that is evidence for (not against) interaction, since the net contribution of all terms involving interaction to var(S), namely var(I) + 2(cov(G+E,I)), may be negative.

How close can we come to an analysis of variance for adoption studies of IQ? The results in Table 1 are rough for several reasons, but illustrative. Define a "−" environment as one in

Table 1 (Dudley). *Environment dominates average IQs in adoption studies.*

		Born into	
		−	+
Raised in	+	112	112
	−	91	NA

which genetically average children raised there will get average IQ scores less than 100, and a "+" environment as an adoptive one or "matched control" (Leahy 1935). There are published studies of children adopted away from "−" environments with IQ data for biological mothers (Skodak & Skeels (1949), nonadopted (half-) siblings (Schiff et al. 1978) or estimated IQ scores for biological relatives (Scarr & Weinberg 1976). Data on children born into "+" environments as compared to children adopted by age 6 months came from Horn et al. (1979) and Leahy (1935). In Scarr & Weinberg (1976) adoption was by age 1 year. Data are apparently not available (NA) for children born into "+" homes and raised in "−" homes, so one cell in the 2 × 2 table is empty. Analysis of variance and tests for interaction are not possible, but one can estimate the main effects if the additive model were assumed.

These data show a strong environmental effect but little or no genetic effect (Horn et al. 1979, p. 196), either by social class of origin (Schiff et al. 1978) or skin color (Scarr & Weinberg 1976). Correlations from some of the same studies suggest genetic effects, due in part to selective placement, or perhaps again showing nonadditive interaction.

Studies of separated one-egg twins are further from allowing an analysis of variance, because the environments of the "separated" twins were often quite similar (Farber 1981; Kamin 1974, pp. 50–62; Newman et al. 1937, pp. 337–41). So environmental *and interaction* terms may contribute to the correlations. In one case environments did make a 24-point difference in "identical" twins' scores (Newman et al. 1937).

The most widely used general method of heritability estimation for IQ is path analysis, assuming the additive model. With interaction, as Wahlsten says, there is no reliable way to estimate the heritability of IQ.

ACKNOWLEDGMENT

This research was partially supported by National Science Foundation Grant DMS-8506638.

One statistician's perspective

Colin Goodall

Program in Statistics and Operations Research, School of Engineering and Applied Science, Princeton University, Princeton, NJ 08544
Electronic mail: *colin@jackknife.princeton.edu*

I wish to compliment the author on addressing the statistical problem of detecting heredity environment interaction so vigorously. Since the target article evokes general issues in statistical modelling and the analysis of variance, to comment fully on it would require at least a short course. Such a course would cover basic 1-way and 2-way ANOVA, the meaning of interactions, and the assumptions underlying the use of analysis of variance. It then would move on to more advanced topics, such as the estimation of contrasts, the interpretation of factors as fixed or random, the analysis of covariance, and more-than-2-factor analysis. In lieu of the course, I recommend an ANOVA classic, Snedecor and Cochran (7th ed., 1980), which includes as examples some agricultural field trials, the same type of data as those mentioned in Wahlsten's paper. The advantage of this text is that it discusses the mechanics of analysis of variance and connects with Wahlsten's paper without the interference of heredity and environment as meaning-laden labels of the factors involved. The mechanics of ANOVA have considerable importance; the *limitations* of ANOVA must be thoroughly understood before heredity versus environment can be debated. Wahlsten appears to agree with this observation: On the surface, his principal point is a statistical one, namely, that large sample sizes are needed to detect interactions.

Moving to newer work, the recent revitalization of the field of analysis of variance has been led by the need for efficient *industrial* experimentation (Box et al. 1978). In the social sciences the text of Fox (1984) is a contemporary exposition of the linear model. The two volumes of Hoaglin et al. (1983; 1985) present the more exploratory aspects of data analysis, including 2-way analysis and transformations, and we can look forward to one or more additional volumes from these editors specifically on analysis of variance. This forthcoming work promises to be an original and perceptive exposition of analysis of variance, and will meet many of the statistical concerns of Wahlsten's paper head on.

Before turning to detailed comments on Wahlsten's paper, it is worthwhile remarking that questions about heredity vs. environment and the use of analysis of variance are much less controversial than C. Spearman's and L. L. Thurstone's development and use of a sophisticated statistical technique, factor analysis, to define an intelligence quotient. Jensen (1985) and commentators provided a general review and discussion, while Gould (1981) explains at length why this is a misuse of statistics. [See also Jensen: "The nature of the black-white difference on various psychometric tests: Spearman's hypothesis" *BBS* 8(2) 1985.] It is fair to say that many statisticians have misgivings about factor analysis as methodology *per se*. On the other hand, whatever the validity and outcome of the heredity versus environment debate, there is no doubt that analysis of variance is one of the most well-founded, best-understood and most reliable of statistical techniques!

An additive relationship is said to exist between a response variable y and two factors α and β if their joint contribution to the response is the sum of a separate contribution from each factor (Emerson & Hoaglin 1983). The additive model, with replication, written

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad (1)$$

is the simplest 2-factor model to fit and to interpret. The change in the mean response, $\mu + \alpha_i + \beta_j$, when the level of the first factor changes from α_i to α_λ, is the same whatever the level of the second factor. In a specific practical situation do we really believe that the additive model is at all realistic? The answer is no: Much more likely than not, there will be some interaction. The appropriate question is, therefore, how large should an interaction be for it to be important for us? To answer this the size of the interaction is compared to the sizes of the main effects and all three are compared to the error variance. (As an extreme case, the interaction may be so large that it is not worth our while separating out one or the other of the main effects.)

Testing an interaction for statistical significance means we strongly wish to avoid rejecting the additive model by chance alone. That is, hypothesis testing tends to give additivity the benefit of the doubt, although we may never really believe it to be true, and we may in fact be prejudiced toward requiring substantial evidence that the interactions are negligible before accepting additivity. If an interaction is to be statistically significant when the underlying population interaction is relatively small relative to the error variance, then the sample must be large. This leads to the classical power computation. In practice we would not, or should not, utilize our resources solely to demonstrate the statistical significance of an interaction that is too small to bother with anyway. Instead, if more data are to be

collected, the model should be refined through the introduction of other variables.

From a somewhat different point of view, suppose that, in a 4 × 4 table, the observed mean-squares for the 2 main effects, the interaction and the error are in the proportions 120:75:20:10. Then the interaction would be statistically significant at a probability of 5% if 6 or more replications were used. On the other hand, with proportions 120:75:15:10, no number of replications would suffice. This example, chosen to parallel the author's Table 2, though admittedly "upside down," illustrates that the interaction must be sufficiently large compared to the error of replication for statistical analysis to give results. In fact, the number of replications matters only if the interaction mean square is between 18.8 and 25.3 (with error mean square 10). For both sets of proportions used by way of illustration we may be perfectly prepared to conclude that the two factors are approximately additive: 120 and 75 are large compared to 15 and 20. An essential caveat to this conclusion is that it applies *within the respective ranges of the two factors* considered. Wahlsten's computation of σ_E, σ_H, σ_I illustrates very clearly how the interaction effects in a model can increase quadratically when the increase in the main effects is only linear.

Broadly stated, the goal of statistical analysis is to find a parsimonious algebraic representation of the given data, with the use of inference tools to determine whether one or more components of the model may be due to chance alone. The choice of model is bound up with the types of departures of the data from the model. Least-squares or nonlinear least-squares techniques (Bates & Watts 1988) are appropriate provided the error distribution is Gaussian. We must therefore ensure that the data in each cell of the 2-way table is Gaussian. A transformation may be required, or separate weights if the variances appear to be unequal (Carroll & Ruppert 1988). Transformation should be regarded, therefore, as an integral part of the statistical analysis.

Emerson and Stoto (1983) point to the serendipitous effects of transformation, a transformation to promote symmetry/Gaussianity may also promote additivity. In specific cases the model is determined by detailed subject-matter specific insight – considerations about the form of a reaction in chemical kinetics, for example. However, questions of heredity versus environment cover so broad a range of subject matter that the model cannot be restricted indiscriminately to untransformed ANOVA. Thus, in general, one or more alternative models should be presented for interpretation in terms of heredity versus environment.

One alternative is a multiplicative model, which is additive on the log scale. Recognizing that the log transformation tends to bring the errors to Gaussianity, it can be fitted by nonlinear or weighted least squares on the original scale, or by ordinary ANOVA on the log scale. Interpretation poses no difficulties, since in percentage terms, the effects of the two factors are additive. If the data are fitted well by a more complex model, such as $Y = aXe^{-bX}$, then this model should hold great intrinsic interest, especially to the theorist of heredity versus environment who wishes to understand *how* the 2 factors X and b interact. The farmer may require a simpler rule of thumb, obtained by linear approximation, and will ignore a small interaction, but that is not a key concern in the theoretical debate.

On the relativity of quantitative genetic variance components

Charles J. Goodnight

Department of Zoology, The University of Vermont, Marsh Life Science Building, Burlington, VT 05405-0086
Electronic mail: *cgoodnig@uvmvm.bitnet*

Narrow sense heritability is defined as V_A/V_P, the ratio of the additive genetic variance, V_A, to the phenotypic variance, V_P (Falconer 1981). Fisher (1958) originally defined the additive genetic variance in terms of the average effects of alleles; however, additive genetic variance plays a central role in quantitative genetic theory because of its functional significance. The additive genetic variance is the component of the phenotypic variance that can contribute to a lasting response to selection. For this reason, narrow sense heritability is important for the field of quantitative genetics because it is directly proportional to the response to natural selection.

The additive genetic variance and heritability of a population are measured by statistical means and they reflect the potential for evolutionary change in a particular population at a particular time. In measuring the additive genetic variance in a population, it is often convenient to divide the phenotypic variance into additive and environmental components:

$$V_P = V_A + V_E.$$

The second component of the phenotypic variance, the "environmental" variance, V_E, needs further exploration. In this partitioning of the phenotypic variance, "environmental" does not mean "nongenetic." For example, only a slightly more detailed genetic analysis is likely the reveal that the environmental variance includes dominance variance, and possibly other forms of genetic variance. In other words, in a quantitative genetic model environmental variance is actually residual variance, that is, phenotypic variance that cannot be assigned an explicit causal explanation.

Broad sense heritability is ideally the ratio of the total genotypic variance to the phenotypic variance. Measuring the genotypic variance, however, will be a difficult, if not impossible, task in a sexually reproducing organism (Falconer 1981). If it can be measured it must still be recognized that the measure of the genotypic variance is for a particular environment and for a particular breeding structure. Thus, a measure of the genotypic variance in a laboratory setting may have little bearing on the genotypic variance in the field.

This does not mean that measures of narrow and broad sense heritabilities are without utility. It simply means that it is necessary to recognize the relativity of quantitative genetic measures. Strictly speaking, quantitative genetic measures are only accurate for the conditions under which they were measured. A properly executed measure of the narrow sense heritability will give an accurate measure of the potential for a response to selection in that particular population in that particular environment. Similarly, broad sense heritability is an indication of the portion of the phenotypic variance that can be assigned to genetic effects in a particular setting. To use an estimate of heritability measured in one population or environment to draw conclusions about heritable variation in another population or environment is an extrapolation and should be recognized as such.

When studying the effects of genetic and environmental influences on the expression of a trait in an organism, a common approach is to use a factorial design experiment that exposes sets of relatives to multiple environments. In this case, an analysis of variance will provide a measure of the resemblance among relatives in the context of this experiment. For example, if sets of individuals with one parent in common are exposed to a range of environments then the variance among these half-sibling families will be a measure of $\frac{1}{4}$ of the additive genetic variance in this set of environments. If selection were to be applied to a population in this set of environments then 4 times the variance among (half) sibling families would be predictive of the response to selection. A problem arises when the factorial experiment does not accurately reflect the population and environment that is being studied. For example, if the environments were chosen to be a range of environments, rather than a random sample of the environments the population naturally experiences, then the experiment is not likely to be predictive of a response to

selection in nature. This will be particularly true if there are genotype environment interactions. As Wahlsten points out, the analysis of variance approach maximizes main effects at the expense of the interaction effects. This is likely to make it difficult to compare such an experiment to a natural population.

If the number of environments is small and the different environments are identifiable, then it is often possible to use multivariate analysis of variance to analyze genetic and environmental factors affecting a population (Via 1984a; 1984b; Via & Lande 1985). As with a univariate approach, a set of relatives are raised in a series of environments. However, in the multivariate approach the genetic variance components are measured separately for each environment and the correlation among the performance of relatives in different environments is also examined. This approach postulates that the expression of a trait in different environments is actually different correlated traits. That is reasonable because for many traits different genes will be expressed in different environments. The power of this approach lies in the fact that there is a separate measure of the additive genetic variance in each environment. The response to selection will then be determined by the responses to selection in each of the individual environments and the genetic correlation among the expression of the traits in the different environments. This approach avoids many of the pitfalls of a univariate analysis of variance because the estimates of the variance components in single environments are not made at the expense of the interaction among environments; however, it is still a linear approximation to what may be a very nonlinear world.

Through the ANOVA looking-glass: Distortions of heredity-environment interactions

Gordon M. Harrington

Department of Psychology, University of Northern Iowa, Cedar Falls, IA 50614-0505

Electronic mail: *harrinni@uiamvs.bitnet*

Wahlsten has forcefully reminded us that statistical manipulations provide a looking-glass that may offer us, as they did Alice, new visions and insights at the risk of distorting reality. Many years ago, I was an expert witness in a federal court hearing on sex discrimination in academe. The defendant's expert presented a total regression analysis (ANOVA) showing that most of the variance in salaries was a main effect of academic rank, sex effects were small, and interactions with sex were negligible. My developmental perspective, to look at factors in the order in which they entered into career development, prevailed; with a stepwise regression first extracting sex effects, then experience, then the interaction of sex with experience, and then rank – sex effects and interactions with sex were large and the contributions of rank were negligible. When factors are correlated, ANOVA will ascribe components of variance based on what best fits the specific data set, including the vagaries which would be relegated to error variance if the study were replicated. Standard textbooks have always emphasized that unless there is a good logical reason to expect the linear equation to represent the situation veridically, "fitting a straight line can be regarded only as an empirical exercise, with no meaning to the constants obtained beyond the purely formal one of specifying the straight line that most nearly represents the observed data" (Ezekiel & Fox 1959, p. 67).

With respect to the salaries, there were interactions. With sex genetically determined and years of experience clearly environmental there was a genetic-environmental interaction. Neither I nor the defense seriously considered gaining additivity by scale transformations. It is hard enough to explain to a judge why we measure variability in square dollars. The prospect of explaining salary differences in log-dollars rather than dollars and variability in square log-dollars is truly daunting – besides which, it is easy to explain that women receive smaller annual increments than men do. I would find it difficult to explain to a judge that the smaller increments are just an apparent effect and the reality is that there is no difference in increments once we measure salaries in log-dollars. For explanatory purposes a model must be logically and theoretically, not empirically, based.

A universal example among those who teach statistics is the observation that the correlation between the annual Swedish birth rate and the number of storks nesting is higher than any other known correlation in the scientific application of statistics. Cited to demonstrate correlation is not causation, it does represent a regression model which fits the data almost perfectly. Hence it offers predictability, the primary requirement for scientific meaning, but it lacks explanatory value. We reject the observed regression as a causal model as a matter of logic relating to matters we understand. In the early history of statistical applications, lacking understanding, psychology was obsessed with empirical curve fitting and every set of data took the investigator to Pearson's (1930) tables to discover which model should be selected for the observed data. But in fact, for *any* set of data, one can empirically fit more than one model perfectly with zero error variance (e.g., a Taylor series or a Fourier series).

Ultimately we learned we needed a logical basis for model selection. Ezekiel and Fox (1959), after fitting several curves to the same data, observed: "It is quite apparent that the differences in the shapes of the several curves are due solely to the particular form of equation used in computing them. There are certain types of relations which can be accurately represented by each of these equations. When it is 'fitted' to data where that type of relation is really present, it can give a curve which accurately represents the central tendency of the data. But when the same equation is fitted to data for which the underlying relationship follows a different function, the resulting curve gives only a distorted representation of the true relation – *it shows the relationship only insofar as it is possible to do so within the limits of the particular equation used*" (emphasis in the original, pp. 101–102).

Suppose, lacking a theoretical model, one chooses to use an empirical approach to interactions to gain additivity. How shall one choose the transformation? Time measures for runway learning in my laboratory certainly need transformation, but should it be log-latency, which fits some strains of rats, or speed, which fits others? These represent two different models of learning. Complete faith in empirical decision making would lead to the possible conclusion that the principles of learning vary with genotype. Powers (1950; 1955) reported crop genetics data which necessitated transformations to handle genetic interactions where different transformations were required for different years of data collection on the same population. Thus, transformations which allowed the data to be interpreted for some purposes sacrificed any possibility of gaining knowledge about other aspects of the data.

One issue of genetic-environmental interactions and power of the analysis of variance has not been addressed. Wahlsten points out that human studies generally report no interaction between heredity and environment. Suppose, however, that the behavior of parents is influenced by the behavior of their children, that is, that parental behaviors are at least in part a response to the behavior of their children. Suppose further that those child behaviors are also subject to some genetic influences. Then, since parental behavior is an environmental influence we have a genetic-environmental interaction or, more precisely, a phenotype-environment interaction. That is, the dependent variable, phenotype, also affects the independent variable, environment. Such phenomena are known to occur in infrahumans where maternal effects may modify genetic effects toward inter-

mediate expression (Fulker 1970). From an evolutionary viewpoint such maternal effects minimize extremes where intermediate behaviors are more adaptive. It is easy to imagine a similar adaptive process in humans – parents increasing their behavior modification activities in response to and counter to behavioral extremes. Then the appropriate model would not be a single function combining heredity and environment additively but a resultant of independent heredity and environment functions with some common variables. If the parental environment is affected by genotype, a regression coefficient of genotype exists in the environmental function that is independent of the comparable coefficient in the genotype function.

It is just such reciprocal interaction conditions which necessitated structural equation modeling in economics (Haavelmo 1943; 1944), an approach with conceptual origins in genetics (Wright 1934). Reciprocal influences have a profound effect on ANOVA. A nonsignificant interaction variance is meaningless as a model of parenting if the interaction is the resultant of two regression coefficients of opposite sign. What is the power of the test? Econometricians would conclude that the ANOVA results are biased rather than unbiased estimates so that increasing sample size cannot increase precision; thus, there is no power. In reviewing the conditions requiring structural equations, Goldberger (1973) points out that this is a substantial rather than a statistical use of the term, *bias*. To be statistically precise, it would be more accurate in the parenting example to say that the analysis estimates only the observations to be expected on replication and does not estimate the underlying parameters or the parameters which would be obtained in a fully controlled experiment. The ubiquity of genetic-environmental interactions in the animal laboratory where experimental control is the rule and their paucity in human studies where control is limited may not reflect a species difference but rather a distorted human image which can be seen in its true form only with theoretically structured models.

Why do gene-environment interactions appear more often in laboratory animal studies than in human behavioral genetic research?

Norman D. Henderson

Department of Psychology, Oberlin College, Oberlin, OH 44074
Electronic mail: *fhenders@oberlin.bitnet*

Wahlsten argues that the prevalence of GxE interactions (GxE) nullifies efforts in human behavior genetics to partition variance into G and E sources, yet this meaningless activity continues because the power to detect GxE is so low. He refers to an "immense collection" of animal studies showing GxE, reviews the long recognized problem of low power to detect interactions, gives an example in which the true relationship between distance, mass, and force of attraction is missed using a factorial ANOVA, and illustrates the low power to detect some multiplicative relationships between G and E.

When Wahlsten (1979) first raised the GxE issue, his selective examples and overstated conclusions moved Fuller (1979, pp. 473) to respond, "Fortunately most persons working in this area are aware of its sensitive aspects, and know the complexities of G-E correlations and interactions." Fuller also questioned the basis for Wahlsten's genetic nihilism, given the data available at that time. The intervening decade has brought a flood of sophisticated large-scale research, continued theoretical attention to GxE (e.g., Henderson 1986; Parsons 1988), and an emphasis on power and design considerations (e.g., Hewitt et al. 1988; Martin et al. 1978) in behavior genetics. Wahlsten has nevertheless restated some of his 1979 argument, missing the point of much of the contemporary human research, as he concentrates largely on the lack of power to detect model-destroying GxE. Fuller's 1979 commentary would apply just as well to this target article.

I shall focus my comments on Wahlsten's "Minneapolis BGA experience," where he heard frequent reports of GxE in animal research but few in the human data.[1] Commentary length limits do not allow me to explore some statistical, philosophical, and motivational issues in the target article. For example: Is Wahlsten's low-power issue as serious as the pervasive tendency in biobehavioral research to ignore GxE by keeping it buried in error terms? Why does the presence of interactions vitiate the partitioning of variation into linear and nonlinear components in random models? What *is* the scientific objective of human behavior genetic research? Why does Wahlsten focus critically on human behavioral genetics (including the eugenics section and the rehash of a long ago discovered typographical error in Jinks & Fulker 1970) when the interaction issue has implications for *all* research areas?

Although I do not characterize the number of animal studies showing GxE as immense, compared with almost none reported in human research, the frequency of GxE does differ between the two research domains. I suspect that this is not because of small-N low power in human versus animal research as Wahlsten implies. The low-power argument pertains to both human and animal research yet we find GxE in the latter, which often involves smaller Ns and less reliable measures than those used in human research.

Wahlsten's contrived example of the ANOVA-oriented physicist provides one clue to the animal-human difference in GxE occurrences. His hapless researcher would probably have flunked Experimental Design 101 for failing to heed an important rule – maximize experimental variance. In a world to be explained in which mass and distance both vary by more than ten orders of magnitude, our experimenter respectively limits the range on these variables to only 4:1 and less than 2:1, much as if he had placed a mask containing a tiny window over the appropriate graph in Wahlsten's Figure 2a and saw little interaction. With the same N but an increasing range of values of mass (m_2) and distance (d), the $m_2 \times d$ interaction not only becomes highly significant, it soon accounts for more experimental variance than the main effects. A log transformation would eliminate the interaction and reveal to our physicist the log m_2–2 log d relationship of Newton's law.

Animal researchers rarely forget the maximization rule; they tend to maximize experimental variance with a vengeance. Variance among a random sample of inbred strains is twice the additive genetic variance found in a random breeding population for a trait. Often strains or selected lines are chosen for their wide differences on a character, resulting in genetic variance many times greater than that of a natural population. Environmental variables (which can be either experimental treatments administered before testing or different test environments) are likewise pressed to maximum practical experimental limits, exceeding normal variation. This animal-laboratory variance maximization strategy is infrequent in human research, where both the G and the E variation of tested samples tend to be equal to or smaller than that of the general population. Returning to the graphs in Figure 2a: If most individuals live in environments falling between levels 1 and 3 and nearly all belong to one of the middle three genotypes, GxE will account for little of the explained variance in all but the $Y = a + bX$ examples. The latter interactions are rare, even in animal research, and may often be artifacts of the limits of the design or the dependent variables used (e.g., Henderson 1968; 1979a; 1979b).

Could we generate more or larger GxE in human research? Yes, but probably not by simply increasing sample sizes. Human studies with large samples often produce estimates of GxE effects sizes near zero and larger samples are not likely to change this finding much. A better strategy might be to mimic some of the animal research paradigms. For example:

1. assemble two genetically different groups of children, one exhibiting moderate mental retardation associated with a single gene and a second group without the deleterious gene;
2. assign half of each group to an environmental condition providing few learning opportunities and the remainder to an "enriched" condition, with training ranging from rudimentary skills and knowledge such as dressing oneself or knowing one's name, to high level skills and knowledge such as playing the violin or solving trigonometry problems;
3. measure these skills and knowledge levels and run a 2×2 ANOVA for each individual variable and several composite scores.

I am confident that this questionable experiment will generate GxE and that these interactions will differ depending on the behavior measured. Large genetic differences observed between the two untrained groups on simple tasks might often be nonexistent in the trained groups. In contrast, both untrained genetic groups would probably score near zero on a trigonometry problem while the two trained genetic groups would probably differ considerably in their ability to solve it. A composite score generated from a mix of easy and difficult tasks (not unlike current psychometric instruments) might show little GxE.

Many variations of this theme exist. Genetic groups could differ on sensory or motor capacity, drug sensitivity, growth rate, etc. and treatment or test environments would be designed to exaggerate or attenuate these differences for highly specific behavior. [See Macphail: "The comparative psychology of intelligence," BBS 10(4) 1987] Although these crude human analogs of laboratory animal research designs are good bets to generate GxE, reports of such effects are largely absent from the literature. Perhaps their absence is not because of low statistical power on an ominous conspiracy among researchers, but because such designs would often be unethical, trivial or uninteresting for the study of human variation. Genotypes could surely be found that respond differentially to environmental influence, but this would hardly vitiate current variance and covariance estimation strategies.

NOTE

1. All six reports of animal GxE at Minnesota involved differential sensitivity of strains or selected lines to pharmacological agents or stress, unusual paradigms in the human research.

A nemesis for heritability estimation

Jerry Hirsch

Departments of Psychology and of Ecology, Ethology and Evolution, University of Illinois at Urbana-Champaign, Champaign, IL 61820
Electronic mail: *jhirsch@h.psych.uiuc.edu*

Congratulations to Wahlsten, who clarifies in fundamental, but elementary, statistical reasoning why the tidal wave (on the positively accelerated "rate of publication," see DeFries & Plomin 1978, p. 473) of unjustified human heritability estimates has done so much harm over two decades to an important research topic. The logic of this methodology is widely misunderstood throughout the scientific community (a particularly influential, and thus all the more unfortunate, example was the comment, "Genetics and hereditable [sic!] IQ" by the editor of the important journal *Nature* propounding his plea: "*Geneticists (and others) should not . . . be fearful* [to] . . . . search for blocks of data that may throw light on the heritability of intelligence"). (Maddox 1984, p. 579, emphasis in original) See Harrington (1988) for a penetrating analysis of "intelligence" testing. The impact of this misunderstanding has been reinforced by a series of reviews in the *Annual Review of Psychology* (Broadhurst et al. 1974; DeFries & Plomin 1978; Fuller 1960; Henderson 1982; Lindzey et al. 1971; Loehlin et al. 1988; McClearn & Meredith 1966), almost universally accepted as authoritative accounts of the growth of scientific knowledge about human behavior genetics. The signal exception is Fuller's (1960, p. 43) first behavior genetics review in that series, which does consider the Anastasi (1958) discussion recognizing the "How much" question as obsolete.

Earlier, without explicitly considering the question of statistical power, for which unique contribution Wahlsten now deserves full credit, several excellent scientists had already analyzed the inadequacy of the heritability solution to the misconceived nature-nurture question (Fisher 1951; Haldane 1946; Hogben 1939; 1951; Loevinger 1943), but their valuable contributions are not considered in the literature to which they are so relevant. It was the inventor of the analysis of variance, R. A. Fisher, who warned against "the so-called coefficient of heritability, which I regard as one of those unfortunate shortcuts which have emerged in biometry for lack of a more thorough analysis of the data" (Fisher 1951, p. 217). Haldane, Hogben, and Loevinger had made explicit the invalidity of generalizations about either heredity or environment because of the existence of interactions. Haldane (1946) showed that in "general m genotypes in n environments generate $(mn)!/m!n!$ kinds of interaction" (Hirsch 1970). In the simplest type of interaction, even

> though the phenotypic response of a genotype might change from one environment to another (therefore the label *interaction*), a similar change would occur for all genotypes. That kind of relationship between genotype and environment would permit generalizations such as "while the impoverished environment depresses all scores, a shift to the enriched environment increases all scores." The majority of possible interactions, however, preclude generalization. For example, Haldane has shown that even in the relatively simple condition with $m = 3$ genotypes in $n = 3$ environments, there are only $(mn)!/(m!)^n n! = 280$ cases out of a total $(mn)!/m!n! = 10{,}080$ possibilities, in which the order of merit of the m genotypes is the same in the n environments, that is, where such generalizations might be permissible. (McGuire & Hirsch 1977, p. 43).

We then went on to show that the "general expression for the proportion of interactions permitting generalization is $1/(m!)^{(n-1)}$. Note how this proportion must diminish as m and n grow large, that is, with each increase in the number of possibilities to be taken into account." (McGuire & Hirsch 1977, p. 43) A minor correction I should make to the opening sentence of Wahlsten's abstract is to add "environment," that is, ". . . attribute . . . to variation in heredity [or environment] only if . . . ," because the restriction on generalization applies equally to both.

Thus, for the human heritability-estimation bandwagon to roll on, it was necessary to assure its target audience, consisting largely of development psychologists and social scientists, that troublesome interactions simply do not exist. Jinks and Fulker (1970) accomplished that mission from the prestigious platform of APA's *Psychological Bulletin*. As Wahlsten correctly calls to our attention, I published Vetta's invalidation of their meaningless test of interaction. Why? Because the *Psychological Bulletin's* editor at that time "refused to allow Dr. Vetta to publish his correction." In his "letter of rejection, he remarked to Dr. Vetta: Your paper does not materially alter the implications of Jinks and Fulker's analysis, at least for a psychological audience." (Hirsch 1981, p. 23) – a sadly correct but revealing commentary on psychology's low status as a science, a state of affairs analogous to that commented on by Meehl (1978, p. 806) in a somewhat different context when he complained that "'soft' areas of psychology lack the cumulative character of scientific knowledge." Fourteen years later, when Peter Schönemann reanalyzed another part of Jinks and Fulker (1970) that dealt with the Shields twin study and thoroughly invalidated their analysis of those important data, once again an editor of the *Psychological Bulletin* refused to publish Schönemann's (1987) important analysis and correction on the false grounds that an

old, no longer relevant or influential method was being criticized, despite the contrary evidence readily available in the *Science,* and *Social Science, Citation Indexes!*

In the *Psychological Bulletin* Plomin et al. (1977, p. 309) could still write: "This truism for the individual [heredity-environment interaction] is simply false for individual differences in a population," and a special issue of *Behavior Genetics* has recently appeared extolling

> The hypothesis testing revolution in human behavior genetics sparked by Jinks and Fulker's seminal paper in 1970 [which] has simply passed by many people in the field. This is not altogether surprising; to a newcomer it is difficult stuff, requiring a reasonable grasp of statistics and polygenic inheritance as well as at least a smattering of calculus and matrix algebra (Martin et al. 1989, p. 5).

At his departure on receiving his degree, Terry McGuire returned my copy of Hogben's *Nature and Nurture* and remarked that "we should have all read this, because he has said almost everything there is to say." So, I terminate this comment with the final words from 50 years ago:

> The application of statistical technique in the study of human inheritance is beset with pitfalls. On the one hand the experimental difficulties of the subject matter necessitate recourse to mathematical refinements which can be dispensed with in animal breeding. On the other there is the danger of concealing assumptions that have no factual basis behind an impressive facade of flawless algebra. The student may recall the words of Wilhelm Ostwald: "Among scientific articles there are to be found not a few wherein the logic and mathematics are faultless but which are for all that worthless, because the assumptions and hypotheses upon which the faultless logic and mathematics rest do not correspond to actuality" (Hogben 1939, p. 121).

How does one apply statistical analysis to our understanding of the development of human relationships

Oscar Kempthorne

CAA Statistical Laboratory, Iowa State University, Ames, IA 50011

The first sentence of the target article's abstract caused me discomfort, which I will try to explain. In fact, it does not make "sense to . . . [etc.]" We have two forces, heredity (H) and environment (E). It is obvious that the result, the outcome (P), is a result of interaction, using this word in a general scientific sense, of these two forces. In addition, there is the obvious possibility of measurement error, which I will denote by M. So we can write P = f(H, E, M). This merely says that P is determined by H, E, M through some formula $f(\cdot,\cdot,\cdot)$. We then hypothesize, and reasonably so, that there are a number of possible heredities and a number of possible environments, and a number of possible measurement errors. Rather obviously we hypothesize that the numbers in each case are very large. We hypothesize that for a given H, say H_i, a given E, say E_j, and a given M, say M_k, there is a measurement P, P_{ijk}, phenotype. It is obvious then that we have a three-factor situation, and we naturally use the ideas of factorial structures, experimental designs and Mendelian genetics. From this, we have a full factorial model:

$$P_{ijk} = \mu + h_i + e_j + (he)_{ij} + m_k + (hm)_{ik} + (em)_{jk} + (hem)_{ijk}.$$

This is a model that is not of full rank. We can adjoin conditions on the parameters $\{h_i\}$, $\{e_j\}$, $\{m_k\}$, and so on, to make the model of full rank. In the conceptual structure we have specified, for definitional purposes, we suppose that some superentity, such as God, knows the whole story. Our task is to obtain data and then to form some ideas about the nature of the "parameters," μ, $\{h_i\}$, $\{e_j\}$, and so on.

One objective that is worth pursuing is to attempt to describe the variation in the attribute P in terms of the classificatory factors H, E, and M. This is what so-called heritability analysis is about. An obvious first step is to suppose that there is no relationship of the variation of *e* to levels of H and E. This may not be true, of course. Another way of expressing this is to assume that there is no interaction *in a statistical sense* of M with H *and* E. We then have

$$P_{ijk} = \mu + h_i + e_j + (he)_{ij} + m_{ijk}$$

This relationship is assumed to hold for all (i, j, k) combinations in the population of (ijk) combinations under consideration.

We now wish to apply this (very near) identity to ideas of heritability. We are interested in understanding the variation of $\{P_{ijk}\}$, which we characterize by variance. By our assumptions

$$\text{var}(P_{ijk}) = \text{var}(\mu + h_i + e_j + (he)_{ij}) + \text{var}(m_{ijk})$$

We can give explicit identification of the terms μ, $\{h_i\}$, $\{e_j\}$, $(he)_{ij}$ by basing them on the *hypothetical* population in which levels of H are equally associated with levels of E. We may assume that the average of $\{m_{ijk}\}$ is zero. So μ is the mean of this hypothetical population. All this is entirely natural and very assumption free.

The trouble starts with $\text{var}\{\mu + h_i + e_j + (he)_{ij}\}$ which equals $\text{var}(h_i) + \text{var}(e_j) + \text{var}((he)_{ij}) + 2\,\text{cov}(h_i, e_j) + 2\,\text{cov}(h_i, (he)_{ij}) + 2\,\text{cov}(e_j, (he)_{ij})$. If all the covariances are zero, life is simple and

$$\text{var}(P) = \text{var}(h) + \text{var}(e) + \text{var}((he)_{ij}) + \text{var}(m)$$

But this is most questionable with respect to any behavioral trait. It is obviously sure that there is association of h_i and e_j. In the context of IQ, one need only look at the various families of the Enlightenment in Great Britain in the last century, the Darwins, the Huxleys. . . .

The problem here described is not mentioned at all by Wahlsten. The next problem, supposing this one has been surmounted, is the role of the interaction contributions, $\{(he)_{ij}\}$. Obviously, there will always be some interaction; we need only think about "some village Hampden, some mute inglorious Milton, some Cromwell" of Gray's Elegy, or a Beethoven born in a slum.

How can we assess the role? A little progress in thought has been made for the hypothetical case in which H and E are equally or proportionally associated. Suppose in our data set a random subset of *m* levels of H is associated with a random set of *n* levels of E and *r* observations are made on each combination. Then there is an ANOVA:

	d.f.	expectation of mean square
levels of H	$m-1$	$\sigma_m^2 + r\sigma_{he}^2 + nr\sigma_h^2$
levels of E	$n-1$	$\sigma_m^2 + r\sigma_{he}^2 + rm\sigma_e^2$
interaction of H & E	$(m-1)(n-1)$	$\sigma_m^2 + r\sigma_{he}^2$
measurement	$mn(r-1)$	σ_m^2

An exposition of this can be found in Kempthorne (1957, Chapter 13) and in various statistical textbooks. The total variance under this model is $\sigma_m^2 + \sigma_{he}^2 + \sigma_m^2 + \sigma_h^2$. The proportion of variance associated with H is

$$\frac{\sigma_h^2}{\sigma_m^2 + \sigma_{he}^2 + \sigma_m^2 + \sigma_h^2}$$

which can, perhaps (but only perhaps) be called "heritability." Then we can see that the sensitivity of the (usual) F-test for interaction depends on the numerator's and denominator's degrees of freedom, (m-1) (n-1) and mn(r-1) respectively. The sensitivity of this test can be obtained from Tang's tables as exposited by Kempthorne (1952, Chapter 11), or in various other works. The sensitivity of the F-test for interaction can be evaluated. It is not at all clear, however, how one can assert that the tests for interaction have much less power than tests for main effects.

This commentator questions the relevance of arguments based on fixed factors, as in Fisher & Mackenzie (1923), and judges much of the discussion of the work of Neyman and of others cited to be irrelevant. It should be recognized that the notions of components of variance and consequent ratios do not have causal content. They are useful in devising schemes of artificial selection. Kempthorne (1978) may be found useful.

It is quite unclear how the ideas of the target article can be applied to human behavior traits: perhaps to human twin situations, though randomness of association of H and E will be absent. The only application in humans is to the study of correlation and regression among relatives. To examine for interaction one could examine the regression of offspring on parent in disparate environments. With an interaction between H and E, the regressions should be different. My main criticism of Wahlsten's paper concerns its failure to distinguish between the nonadditivity of H and E and association of H and E in the data under study. It is suggested that most of the literature on heritability in species that cannot be experimentally manipulated, for example, in mating, should be ignored.

This commentator agrees emphatically with the last sentence of the abstract.

Heredity and environment: How important is the interaction?

Paul Kline

Department of Psychology, University of Exeter, Exeter, Devon EX4 4QG, England

Electronic mail: *kline@exeter.ac.uk*

It must be made clear, from the outset, that this commentary is written by, to quote Wahlsten: One who knows the field rather than the power calculations of analysis of variance. Yet from this perspective there are several points above this target article that need to be made.

Wahlsten's main argument, that analysis of variance is relatively insensitive in the detection of interactions with the result that heritability estimates are not appropriate, at least in their present form, seems well taken. However, his implications and conclusions may not be so readily accepted.

The first difficulty concerns the psychological meaning. All workers in the field accept that psychological traits or characteristics are influenced by both genetic and environmental factors and that genetic factors operate necessarily within an environment. Thus, in this sense the importance of interaction is a given. As Wahlsten argues, this does not invalidate an additive model. Even more important, the claim that there is a correlation between genetic and environmental factors does not preclude additivity: e.g., in his example "when a bright child is given advantages." This example could easily be extended so that we can think of a bright child actually creating a more stimulating environment by going to libraries or joining in adult activities. In fact, this is a model which many developmental psychologists would regard as quite reasonable, both in respect of intelligence and other traits of ability and personality. Because it is consonant with the partitioning of variance into the main factors and with an additive model, the fact that analysis of variance is weak in the detection of interactions does not seem important.

Other arguments are raised by this example. The first is a simple logical issue: The fact that analysis of variance does underestimate interactions does not mean that interactions must be present and that, therefore, biometric analysis must be worthless, as is implied in the target article. Indeed, as the example of the previous paragraph shows, this is far from the case. The second point concerns the general overstatement of the case. For example, Wahlsten argues in the case of gravity that it makes no sense to claim that a person's weight depends more on body size than planet of residence. Surely this cannot be so since in calculating a person's weight, in any place in the universe, the body size is constant, and all we need to know is the mass of the relevant planet.

Similarly, even if there is a genuine statistical interaction it does not mean that it is nonsense to attempt to quantify the size of the contribution of the two main factors. It could well be the case that the environment contributed far more than did genetic factors to the variance, although there was also an interaction. What Wahlsten has shown is that the contribution cannot be calculated with a simple additive model rather than that the computation is not meaningful. Certainly it would be useful for theory building to know the relative contributions.

There is a further aspect of Wahlsten's argument that I believe to be unsound. This concerns his use of nonhuman examples. With the possible exception of the higher apes (and here the effects would be small), man alone strongly influences the environment in which he lives. Thus, experiments in which animals are placed into static environments (and indeed all studies with animals) are barely analogous to the human case and extrapolation is dubious.

A final, more general, point needs to be made about this paper. This concerns the use of results of research where the contributions to the variance are computed with additive models. If theories are constructed from the results that preclude interaction then the warnings of Wahlsten's paper are serious, but it is doubtful that they are. First, as was discussed, some meanings of interaction imply only correlated genetic and environmental determinants and second, whether interactions are present or not, research into this subject is bound to involve both determinants, and interactions, if present, will be discovered that way.

It must be concluded, therefore, that Wahlsten's target article is valuable especially for users of biometric analysis, because it alerts them to a potential source of error. In practice, however, it appears that this source of error is unlikely to lead to poor interpretations except where study of the problem is left entirely to biometric methods.

Flechsig's rule and quantitative behavior genetics

H.-P. Lipp

Institute of Anatomy, University of Zürich, CH-8057 Zürich, Switzerland

Wahlsten's target article is lucid, polite, and subversive. He uses the genetic field's own weapons to shake the main conceptual pillars of classical behavior genetics. His most effective tool is the analysis (even if a little too pat) of gravitation by means of ANOVA. I am, by and large, a partisan of Wahlsten's view and thus enjoy his bold move. His article may not convince proponents of the line of research under attack, but it is sure to sow a seed of doubt among nonpartisan scientists about the applicability of linear statistical methods to dynamic processes underlying brain development and the organization of behavior.

Much of my own work is based on the use of genetically defined animal models to study the covariation of brain traits and behavior, yet from the viewpoint of individual development of the brain (Lipp et al. 1989). This has led to a conclusion similar to Wahlsten's, namely, that linear least square methods for partitioning hereditary and environmental sources of behavioral variation are inappropriate tools for addressing psychobiological problems concerned with the individual (the focus of interest of most psychologists), and may easily lead to false conclusions with regard to populations (the proper domain of behavior genetics). This judgment is not based on primarily statistical arguments; it is derived from the nature of the behavioral phenotype which is dealt with by behavior genetics using linear

least square methods. The issue has been addressed by Wahlsten briefly (section 14, p. 19).

Let us assume an allele with specifically behavioral effects, a "psychogene," and let us assume that the adult brain represents a hierarchy of dynamically interacting systems; let us add to this what is known about the dynamics of mammalian brain development (Lipp 1979; Lipp & Schwegler 1982; Lipp 1989). The first problem that arises concerns how a "psychogene" can achieve reliable observability at the behavioral level at all. Its expression and penetrance is hindered by a formidable array of buffering and correcting mechanisms. These include system homeostasis, behavioral adaptation by means of learning, adult plasticity of the brain, and a battery of developmental mechanisms for cerebral reorganization such as axonal sprouting and regression, cell death, and other processes of neuronal competition, collectively labelled developmental buffers (Finlay et al. 1987; Katz & Lasek 1978; Katz 1982; Lipp 1979). Each of these mechanisms is capable of masking the behavioral consequences of a psychogene, and none of them obey the rules of linear interactions and additivity required to assess properly the contributions of heredity and environment. Moreover, they are sensitive to other genetic influences and to a host of environmental factors. [See also Johnston: "Developmental Explanation" *BBS* 11(4) 1988.]

Nevertheless, psychogenes probably do exist. Otherwise, rapid selective breeding for behavioral traits would be impossible (Bignami & Bovet 1965; Bovet et al. 1969; Collins 1979; DeFries et al. 1970; Masur & Benedito 1974; Van Abeelen et al. 1973; Van Oortmerssen & Bakker 1981). It would seem that the reliable expression of behaviorally relevant alleles reflects the operation of the *Flechsig's rule* (the "myelogenetic law"), according to which hierarchically superimposed brain systems develop late – a criterion used to define cortical regions with protracted myelination such as associational areas (Flechsig 1920). Hence, psychological specificity of genes can simply be encoded through timing – the latest acting genes automatically affecting the top-ranking system levels of the brain which are presumably responsible for managing the overt behavior of an individual. A familiar example is given by the set of genes that control the onset of puberty. In terms of complexity, the resulting changes in brain and behavior triggered by the rising hormone levels most certainly exceed other somatic changes. Yet the causative gene action is relatively simple. In general terms, the behavioral consequences of many genes are not a property encapsulated in the DNA, but depend on the developmental stage and configuration of the target systems in the brain.

Accepting Flechsig's rule leads to the paradoxical conclusion that single gene effects on behavior are most likely to be observed by concentrating on *complex* behavioral traits: If a gene acts very late, its effects escape the extremely powerful masking effects of developmental reorganization and are modified solely by the system homeostasis of the top-ranking systems, or by learned behavioral adaptation. Thus, its effects are more easily discovered than those of earlier acting alleles, be this by an observer or by natural selection. On the other hand, the behavioral expression of such an allele is inevitably variable, strongly dependent on particular environments, and prone to interaction with both environment and genes. Moreover, if late-maturing brain systems are the chief targets of "psychogenes," it also means that these systems remain modifiable by environmental influences during a long ontogenetic period: To the behaviorally relevant brain systems, it matters little whether it has been modified by a gene or by an external factor (Lipp et al. 1988). Thus, depending on the conditions, the very same allele may at one time result in a sample of phenotypes with a high degree of heritability, and another time in a sample characterized by much environmentally dependent variability. It would seem that the ongoing debate and conflicting results on the heritability of schizophrenia (Byerley et al. 1989; Kennedy et al. 1988; McGue et al. 1983) may in part reflect such interaction between gene effects and developmental properties of the brain (Lyon et al. 1989). If this scenario is correct, there is obviously no logical way to disentangle hereditary and environmental contributions to a particular behavioral phenotype, and the results of the classical partitioning by means of ANOVA are meaningless, certainly for predicting the behavior of an individual, but perhaps also for theories of population genetics and evolution. For example, selective breeding for behavioral extremes might well succeed for a trait that has not shown much heritability.

In terms of research strategy, this line of thought reaches the same conclusions as Wahlsten: To assess the effects of genes on behavior, it is first necessary to understand how variations of the brain develop and how they influence behavior. Understanding the target paves the way for understanding how genes modify it. This is not to denounce the computational attempts of behavior genetics as generally meaningless, for if strong heritability of a phenotype is found, it is indicative of gene action on brain and behavior, and thus heuristically meaningful. The underlying theory may be wrong, but the results remain valuable for the interested neurobiologist.

ACKNOWLEDGMENT

This paper was supported by SNF 3.206.-88.

Why are interactions so difficult to detect?

Scott E. Maxwell

Department of Psychology, University of Notre Dame, Notre Dame, IN 46556

Electronic mail: *srggcc@irishmvs.bitnet*

Wahlsten has provided a valuable message not only to researchers in behavior genetics but also more broadly to behavioral researchers in general. Although, as he acknowledges, at least some statisticians have been aware for several decades of problems of low power for testing interactions, applied behavioral researchers have by and large shown no sensitivity to this phenomenon. In particular, behavioral researchers continue to seek interactions in factorial designs with very small cell sizes, apparently oblivious to the likely impact of low power on their results. This practice has continued despite warnings from Cohen (1977, p. 375) that interactions are typically more difficult to detect than main effects. Similarly, more than a decade ago, Cronbach and Snow (1977) provided explicit sample size advice for behavioral researchers interested in detecting an aptitude-treatment interaction (ATI). They stated: "An ATI study is an experiment, and most investigators have followed experimental tradition, employing 40 or fewer Ss per treatment. This is radically wrong. The ATI study must be much larger than a study where main effects or single correlations are at issue" (1977, p. 46). Wahlsten is right that despite such admonitions from recognized methodological authorities, most behavioral investigations of interactions are extremely susceptible to Type II errors. Indeed, Sedlmeier and Gigerenzer's (1989) recent survey of the power of studies published in *Journal of Abnormal Psychology* suggests that typical power values in this area are even lower than Cohen (1962) found in a comparable survey twenty-five years earlier.

Although the reasons for continued inadequate power are certainly complex, one contributing factor in the case of interaction research may be the lack of understanding on the part of most researchers as to why tests of interaction often lack power. In particular, researchers may have failed to appreciate that interaction tests frequently lack power for two reasons, one intrinsic to behavioral phenomena, and the other often reflecting a suboptimal method of data analysis.

One reason that interaction tests may suffer from low power is that the effect size for the interaction (f_I) may be smaller than the

effect size for a main effect (f_H and f_E). Although it is mathematically possible for f_I to far exceed f_H and f_E, in reality it often happens that f_I is smaller than f_H or f_E, or both. For example, in every situation simulated by Wahlsten, f_I is smaller than either f_H or f_E (see his Table 4). Such an outcome is to be expected when the interaction is ordinal with respect to the main effect factor in question.

To consider this point in detail, consider a 2×2 design where μ_{jk} is the population mean for the cell in row j and column k. Without loss of generality, assume that the rows are arranged so that $\mu_{11} < \mu_{21}$. The interaction is ordinal with respect to the row factor if and only if $\mu_{12} < \mu_{22}$. Given the usual side conditions that effect parameters be constrained to sum to zero, cell means can be written as

$$\mu_{11} = \mu - \alpha - \beta - \alpha\beta$$
$$\mu_{12} = \mu - \alpha + \beta + \alpha\beta$$
$$\mu_{21} = \mu + \alpha - \beta + \alpha\beta$$
$$\mu_{22} = \mu + \alpha + \beta - \alpha\beta.$$

However, $\mu_{11} < \mu_{21}$ implies that $-\alpha < \alpha\beta$. Similarly, $\mu_{12} < \mu_{22}$ implies that $\alpha\beta < \alpha$. Together, then, $-\alpha < \alpha\beta < \alpha$, so that an ordinal interaction implies that $(\alpha\beta)^2 < (\alpha)^2$. As a consequence, the effect size for the interaction will be less than the effect size for the row main effect. Such a pattern will occur when one environment (or one strain) is optimal for everyone, but the precise magnitude of the advantage of this environment (or strain) over others varies as a function of strain (or environment). To the extent that environments and strains operate in this fashion, interaction effect sizes will necessarily be smaller than at least one of the effect sizes for the main effects.

Even if the interaction effect size were equal to the effect size for the main effect, the power for the interaction test might be lower than the power of the main effect test, because power also depends on effective sample size. The ϕ index for an effect in a factorial design is approximately equal to the product of f and n', where n' is given by

$$n' = \frac{N - ab}{df_{effect} + 1} + 1.$$

When both factors have more than 2 levels, the degrees of freedom for the interaction effect will exceed the degrees of freedom for either main effect. As a consequence, n' will be smaller for the interaction than for either main effect, and power will suffer accordingly.

Thus, tests of interactions often suffer from two disadvantages relative to tests of main effects. While a lower effect size is often an intrinsic disadvantage, the second disadvantage of larger df_{effect} is to some extent under the experimenter's control. This second disadvantage occurs because the interaction test is a global one, which is somewhat sensitive to all forms of interaction but may not be highly sensitive to any particular form. As Rosenthal and Rosnow (1985) and others have pointed out, however, planned comparisons of interaction contrasts may increase power substantially. For example, Levin (1975) shows that ϕ for a planned contrasts is given by

$$\phi = \sqrt{n\psi^2/2\sigma^2\Sigma c_j^2}.$$

It can be shown that ϕ for the optimal interaction subeffect can be written as

$$\phi = \sqrt{n\Sigma\Sigma(\alpha\beta_{jk})^2/2\sigma^2}.$$

In contrast, ϕ for the global interaction effect is given by

$$\phi = \sqrt{n\Sigma\Sigma(\alpha\beta_{jk})^2/[(a-1)(b-1)+1]\sigma^2}.$$

When both factors have more than 2 levels (i.e., $a > 2$, $b > 2$), ϕ for the planned comparison can greatly exceed ϕ for the interaction. For example, in Wahlsten's "Y = a + bX, Case 1" example, ϕ for the optimal interaction subeffect is larger than ϕ for the H main effect. A test of this specific interaction hypothesis may yield more power than the test of the H main effect for this configuration of cell means. Thus, to the extent that the experimenter is able to anticipate correctly the pattern of cell means, power to detect nonadditivity can be greatly increased by testing planned comparisons instead of the global interaction.

Some researchers might argue that nonsignificant interactions are not problematic, since the interaction sum of squares simply serves to lower proportions of variance accounted for by main effects. Wahlsten convincingly argues, however, that developing a correct model of the phenomenon under study is more important than partitioning variance. In this respect, the traditional variance partitioning approach underlying ANOVA has obscured the goal of developing a model to understand behavior. This goal is consistent with recent developments in behavioral statistics texts (e.g., Judd & McClelland 1989; Kenny 1987; Maxwell & Delaney 1989) toward approaches that emphasize building models and comparing them not just for multiple regression but for ANOVA as well.

In summary, Wahlsten is right that interactions are often difficult to detect. The fault lies not with the statistical techniques themselves, however, but with the intrinsic nature of many behavioral interactions and the frequent use of suboptimal data analysis techniques. A priori power calculations are essential. Fortunately, power calculations are becoming feasible even for complex designs with the development of appropriate software. Finally, researchers are encouraged to formulate specific interaction hypotheses when possible, and to test corresponding planned comparisons.

Who believes in estimating heritability as an end in itself?

Peter McGuffin[a] and Randy Katz[b]

[a]University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, United Kingdom and [b]Department of Psychology, Toronto General Hospital and University of Toronto, Toronto, Ontario, Canada M5G 2C4

Wahlsten has provided a perceptive, lucid, and well-argued critique of some of the shortcomings of "heritability" coefficients, focussing in particular on the hazards of overlooking gene-environment interactions. However, we wonder if, in emphasizing some of the more blatant misuses of the concept of heritability, he has effectively set up a straw man for enthusiastic dismemberment. For example, we think it would be a hard task to find a reputable human behaviour geneticist who *disagreed* with Wahlsten's strictures on the misuse of heritability by would-be eugeneticists. Furthermore, we would not even attempt to defend the idea that estimation of heritability can be seen as an end in itself. Indeed, Wahlsten's final sentence quoting Bateson (1987), is very similar to the views we have previously expressed (McGuffin & Katz 1986) that "calculation of heritability is in itself an empty exercise if it does not lead to a more specific consideration of the ways in which genes and environment coact and interact to produce the phenotype."

Quite apart from the problem of arriving at an inflated estimate of additive genetic effects at the expense of multiplicative gene-environment effects, estimates of heritability have other short-comings which Wahlsten does not mention. Commonly applied models often require the prior assumption that nonadditive *genetic* factors are negligible, i.e., that there is no dominance or epistasis. At an even more fundamental level it is important to remind *BBS* readers, particularly nongeneticists and nonbiometricians, that heritability is the proportion of variance accounted for in the *population*. It does not have a simple meaning at the individual level. Thus, suppose we find that the heritability for IQ is .5 in a certain population. We cannot then take an individual member of that population and

declare that half of the IQ score is determined by his genes. Similarly, heritability is specific to the population in which it is estimated. Other populations may differ with respect to genetic or environmental variance or both and hence the ratio of genetic to total phenotypic variance may differ too. Despite all of these caveats, we believe that the concept of heritability and more generally the practice of partitioning components of the phenotypic variance, does have some utility. Our own work has focussed mainly on the genetics of abnormal behaviour and so we will take three examples from this area to illustrate our case that estimation of variance components can provide useful insights and pointers for further research.

1. Heritability as a pointer to diagnostic validity. The introduction of operational definitions of mental illness (e.g., American Psychiatric Association 1980) has overcome most of the earlier problems of reliability for researchers in this field. The question of diagnostic validity, however, still presents serious difficulties. For example, "schizophrenia" could now be defined in a variety of ways, all of which have acceptable reliability (Kendell 1982). When multiple definitions have been used in practice in the same study, however, they have been found to overlap only to a modest extent (Brockington et al. 1978). The researcher, therefore, has a real dilemma about which definition to use. One possible solution is to use aetiological factors to select a definition which is valid as well as reliable. Probably the best aetiological clue in schizophrenia derives from the consistent evidence from genetic studies (Gottesman & Shields 1982), which suggests that the heritability is substantial (McGue et al. 1985). Blindly rediagnosing a series of schizophrenic probands and their monozygotic or dizygotic co-twins using a variety of reliable operational definitions of schizophrenia led to the conclusion that not all of these delineate an equally heritable syndrome (Farmer et al. 1987; McGuffin et al. 1984). Some definitions, for example the one embodied in the diagnostic and statistical manual, 3rd edition (American Psychiatric Association, 1980), provide a heritability of around 80%, but when the disorder is defined using only Schneider's (1959) first rank symptoms, the heritability is effectively zero. The results provide a strong hint as to which definition to choose if the focus of research is to be on biological or genetic aspects of schizophrenia.

2. Heritability, family environment, and eating attitudes. The aetiology of anorexia nervosa (AN) is even more obscure than that of schizophrenia. However, AN is known to respond to family psychotherapy, and family psychodynamic factors have often been invoked as partial causes of the condition. Recent evidence also suggests, however, a genetic contribution (Holland et al. 1988). To investigate these further, Rutherford and her colleagues (in preparation) analysed Eating Attitude Tests (EAT) scores (Garner & Garfinkel 1979) in a twin sample of normal young women. High EAT scores are said to detect cases of AN with good sensitivity and specificity and so studying the sources of variation in normal EAT scores is of potential interest to the clinician. Like Wahlsten, we generally favour the use of likelihood ratio tests in comparing models. We therefore used a computer program to compute the maximum likelihood of a general model in which phenotypic variance was the sum of additive genetic variance (VG), variance due to common family environment (VCE) and the variance due to "specific" environment not shared within the family (VSE). This was compared with reduced models in which each of the variance components in turn was fixed at zero. We found that although the major source of variation in EAT scores (about 60%), was environmental, this consisted entirely of VSE, with VG accounting for the remaining 40%. In the full model, with no bounds set on the parameter estimates, VCE tended to become slightly negative, suggesting that, if anything, family environment makes young women less rather than more like their siblings. These findings may be surprising for some psychodynamic theorists, who emphasize explicitly intrafamilial factors, but are similar to the results of studies on other personality traits (Henderson 1982). We suggest that studies of this type can provide important preliminary information for attempts to investigate environmental aetiology more directly, as well as an antedote to the adoption of a priori assumptions concerning sources of environmental influences.

3. Heritability estimates and clues to the causes of depression One of the problems about aetiological studies in mental illness is that researchers have always tended to approach potential causal factors one at a time. A review of the literature of depression (McGuffin & Katz 1986), however, suggests a complex range of phenotypes, all tending to be familial but having variable heritability and common environmental components. We therefore mounted a study of life events and other forms of adversity in a series of probands with clear-cut onsets of depression presenting to the Maudsley Hospital, London. A number of unexpected findings emerged (Bebbington et al. 1988; McGuffin et al. 1988). For example, not only was depression more common in the relatives of depressed probands than in the general population but so also were threatening life events. Furthermore, although the general population sample showed a very strong relationship between recent threatening events and depressive disorder at the time of interview (Bebbington et al. 1981), there was only a modest difference in the proportion of current "cases" of depression between those relatives who had or had not experienced recent threatening events. Thus, when we applied a logistic regression model with presence or absence of current depression as the dependent variable, both recent life events and relationship to a depressed proband proved to be highly significant explanatory variables but there was also a highly significant interaction effect between life events and family membership. Finding that both liability to depression and propensity to experience life events as threatening are familial and interact raises the possibility that event-associated depression is something which occurs in hazard-prone rather than just stress-susceptible individuals. Although these complex results are clearly not compatible with a simple additive model, such a model provided the essential starting point for our investigation.

In conclusion, we agree with most of Wahlsten's criticism of heritability coefficients and agree with him that human behaviours are often likely to be more causally complex than the usual representation in simple additive models. We remain convinced, however, that simple models provide the best place to begin. Estimation of variance components can provide valuable insights provided they are not awarded undue respect. Thus, it is of no interest whatsoever to know that the heritability of IQ is "really" .8 rather than .6 or .4 but it is potentially of considerable value to know that both genes and family environment make a significant and substantial contribution to the variance of IQ or proneness to depression. Discovering this is a necessary first step, but it is just a first step, and few would argue that it is sufficient in itself.

Good, bad, and ugly questions about heredity

Helmuth Nyborg

International Research Center for Psychoneuroendocrinology, Institute of Psychology, University of Aarhus, DK-8240 Risskov, Denmark
Electronic mail: *psynbo@dkarhø2.bitnet*

Models of nature-nurture interaction date back at least to the golden age of ancient Greece. The models have not grown sufficiently in sophistication, however, to accommodate recent findings, so a more dynamic, nonlinear approach is now called for.

Plato represented a moderate environmentalist position: Men are born noble but society can easily corrupt them. His

colleague, Aristotle, was a convinced hereditarian: People are born different and the differences can and should be exploited by society. The medieval church claimed: Man is born with original sin, but a good Christian upbringing helps. Rousseau declared that we are savages born noble, but that we are corrupted by a less than perfect society. In early nature-nurture debate, the attitude of the participants *was* the evidence, exceptions went largely unnoticed, the nature of hereditary and environmental variables was left unspecified and the assumptions of independence and linear relationships were not tested. Nobility, sinfulness, and corruptibility eventually dropped out of the nature-nurture vocabulary because they were too elusive.

Other abstract variables took their place, however. The contemporary idea of heredity reflects a coefficient based on individual differences around a population mean. Interaction means statistical interaction. Statistics is used to determine how much of average development and functioning is due to the stabilizing (additive?) effects of genes and how much is due to (additive?) modifying effects of the environment. Gene effects are typically assumed rather than localized and specified. Environmental effects refer to social or cultural dimensions intuitively deemed important by the investigator.

One is left with the impression that the nature-nurture debate still operates at a very high level of abstraction and intuition. There is nothing inherently wrong with abstraction or calling things by names, but there must be a certain reality behind it. I suggest that the recent tremendous progress in molecular biology may help us better discriminate between fact and fantasy and distinguish the good, the bad, and the ugly questions about nature-nurture interaction.

From this point of departure, good questions are:

What are the chemical characteristics of the particular DNA material the fetus received from its father and mother?

What structural and chemical developmental effects did this particular combination of DNA give rise to bodywise and brainwise?

To what extent, through which mechanisms, and in which ways are the combined DNA actions influenced by well-defined physical and chemical influences of the environment and vice versa?

These are good questions because the variables can be operationalized and studied by the powerful tools of the natural sciences.

The question: "Why not assume that heredity and environment reflect well-defined, independent, and linearly related variables?" was not originally a bad question, because it led to preliminary evidence that genes count in development. But, it is ugly to continue on this track, as we now know that the idea of independence and additivity no longer holds, as illustrated so well by Wahlsten. The major question: "What is the relative contribution of heredity and environment in explaining the total phenotypic variability for a given trait?" is really a bad one, because (1) linear models obscure the existence of dynamic interactive relationships between heredity and environment, (2) statistical solutions are not likely to settle this problem, and (3) modern developmental biology now acknowledges the importance of nonadditive processes (Pritchard 1986).

Nonlinear models, however, may be quite difficult to apply. The adoption of such models by the natural sciences has led to much controversy, and parts of modern physics have become "entangled" (Glashow 1988). Can we expect similar chaos in the behavioral sciences after having docked linear nature-nurture models? Not necessarily. Dynamic nonlinear models for variable expression of genes have already been developed for the area of neuroendocrinology (Nyborg 1983; 1984) and seem able to explain rather complex aspects of the dynamic biphasic relationships between genes, hormones, body and brain development, functioning, and behavior (Nyborg 1988; 1989; submitted a; b; Nyborg & Boeggild 1989). Briefly, these models reflect the observation that a microscopic dose of sex hormone can selectively enhance or suppress the protein production of thousands of genes, with cascades of early organizational and later activational effects on the development and functioning of body and brain and, accordingly, on behavior. "Optimal" development and functioning seems to depend on intermediate plasma sex hormone concentrations. Lower and higher plasma concentrations both have detrimental effects although for different reasons. Further process nonlinearity arises because the actions of sex hormones are highly sensitive to certain changes in environmental conditions and to the considerable variation in receptor availability, sex hormone binding globulins, and turnover rate. All this speaks for treating each individual as a self-contained dynamic system of processes that interacts with its surroundings in nonlinear ways but within limits set by its DNA material, by ongoing physiological processes, and by the character of environmentally modulated changes in neurotransmitters, including those caused by other people. Temporary high-level stress in a pregnant woman may, for example, permanently switch myriads of fetal genes on or off, with profound long term effects on development and functioning.

All this illustrates what is wrong with traditional models. Assumptions of independence and additivity are often violated, and they lack precision about the character, mechanisms, and locus of action of the relevant causal variables. They group individuals by the thousands, use averaging population statistics, and then make implicit inferences about individuals. The new model will focus first on the chemical and physical agents and physico-chemical processes that make each individual different and then look for communalities (Nyborg 1977, 1987).

ACKNOWLEDGMENT

This work was supported by grant M12-8300 and M15-6870 from the Medical Research Council and from the Research Council for the Humanities, Denmark.

*Present address: Laboratory of Neuroendocrinology, Rockefeller University, New York, NY 10021-6399.

Trying to shoot the messenger for his message

Robert Plomin

Center for Developmental and Health Genetics, College of Health and Human Development, The Pennsylvania State University, University Park, PA 16802

Electronic mail: *pvq@psuvm.bitnet*

The target article is useful in emphasizing the need to study genotype-environment interaction and the large samples required to do the job. However, there is not much new here except for two conclusions, and these are wrong. One conclusion is that the presence of genotype-environment interaction (GxE) "renders a heritability coefficient meaningless." The other conclusion is that difficulties in finding GxE using the traditional analysis-of-variance model means that we should abandon this model. Using the traditional model, the message from the literature is that it is difficult to demonstrate GxE; Wahlsten's remedy is to try to shoot the messenger.

Biased reporting of the GxE literature. Readers of the target article may not be aware that this GxE ground is well worn. The biased reporting of the GxE literature implies that behavioral geneticists have not considered GxE. For example, the target article states that "Plomin's (1988) view" is "that H and E are additive" and that behavioral genetics "'is only useful' for partitioning variance." The fuller quote is that "according to Plomin (1988): Behavioral genetics is only useful for addressing the extent to which genetic and environmental variation contribute to phenotypic variation in a population." The target article misuses this quote to imply that I have not considered GxE. The quotation was taken out of context from a discussion of a completely different topic: "*Unless you are interested in individual differences within a species you will not be interested in behavioral genetics* because behavioral genetics is only useful for addressing the extent to which genetic and environmental variation contribute to phenotypic variation in a population" (Plomin 1988, p. 107).

Of the many papers and books I have written relevant to the topic of GxE, why does the target article use a six-page response to a book review (Plomin 1988) as the representative of "Plomin's view?" Contrary to the impression given by the target article, I have tried hard, but in vain, to find GxE. For example, a 1977 *Psychological Bulletin* paper entitled *Genotype-environment interaction and correlation in the analysis of human behavior* (Plomin, DeFries, & Loehlin 1977) considered how GxE affects twin and adoption estimates of genetic influence and proposed a new test of specific GxE using adoption data, a test I used subsequently in research on GxE. In addition, our text on behavioral genetics (Plomin, DeFries, & McClearn 1980; 1989) discusses GxE and two recent books include chapters on GxE (Plomin 1986; Plomin, DeFries, & Fulker 1988). The 1988 book presents extensive GxE analyses in infancy and early childhood using data from the longitudinal Colorado Adoption Project. Little evidence is found for GxE when the most widely used measures of environment and development are employed. This chapter is also relevant to the target article's implication that behavioral geneticists are unaware of the relatively low power of tests of GxE:

> Limitations on the capacity of the CAP sample to detect interactions should be mentioned. As always, our results are bounded by sample size, for instance. The probability of detecting significant interactions will increase as the number of subjects increases, as the number of variables decreases, and as the amount of variance explained by the interaction increases in proportion to the total variance explained by the multiple regression (Cohen & Cohen 1983). Given the CAP sample size and R^2 of 10 to 20%, our analyses had approximately 80% power to detect interactions that account for 5% of the total variance. However, if interaction effects account for as little as 1% of the variance, one would need a sample size of more than 600 to detect a significant interaction with 80% power given an R^2 of 10 to 20%. One could argue that interactions that account for less than 1% of the variance are not very important (Plomin et al. 1988, p. 250)

In summary, though I stand accused of ignoring GxE, I have in fact developed methods and collected extensive data in an attempt to identify specific GxE.

Animal studies. The target article states that in contrast to human research, "an immense collection of well-controlled laboratory studies of animals has provided abundant evidence of significant and illuminating interactions between heredity and environment." I disagree. The most systematic research on the topic was conducted by Henderson (1967; 1970; 1972) in a series of studies involving thousands of mice. In one study, for example, (Henderson 1972), mice from six inbred strains and their hybrid crosses were reared in impoverished or enriched conditions for the first six weeks of life. As in hundreds of other strain studies, escape-learning proved to be substantially influenced by genetic factors as evidenced by large strain differences. However, rearing environment and the interaction between genotype and rearing environment had little effect. Henderson's earlier work and dozens of other studies occasionally report significant genotype-environment interactions. However, the significant interactions are not consistent within studies, nor do they replicate across studies. Moreover, the interactions, although significant given the large samples used in these studies, account for minuscule portions of variance. For example, a study of 12 Drosophila strains reared under 20 different environmental conditions (Taylor and Condra 1978) found several significant GxE interactions, but, as noted by DeFries (1979), the largest effect only accounted for 2% of the total variance.

Heritability and GxE. The target article is wrong in concluding that the presence of GxE "renders a heritability coefficient meaningless." Main effects and interactions are independent – main effects of G and E are not invalidated by the presence of GxE interactions (Plomin et al. 1980; 1989). In the target article's example of different inbred strains of mice reared in different environments, finding an interaction between strains and environments has no effect on the main effects of strains or of environments.

The target article seems more concerned with denigrating heritability than with finding evidence for GxE. The last sentence of the abstract, for example, states that if the calculation of "'heritability' coefficients is abandoned, interactive relationships can be examined more seriously and can enhance our understanding of the ways living things develop." What is preventing Wahlsten from doing this if he so chooses thus demonstrating to the rest of us how a more serious examination of interactions can enhance our understanding of the ways living things develop?

Does Wahlsten really believe in a model that says that behavior is solely a function of the interaction between G and E? This position carries the implication that there are no "main effects" of the environment. That is, a completely interactive model means that environmental effects cannot be isolated because they are hopelessly enmeshed with the effects of heredity.

If Wahlsten really believes in a purely interactive model, why does he not specify what the expectations of his model are for such designs as MZ and DZ twins reared apart and reared together as well as for other family and adoption designs to show that his model fits the data better than the traditional analysis of variance model? There is no conspiracy against interaction: If an interactive model could be shown to fit the data better than the traditional model, researchers would be quick to use it.

In summary, it is a lot easier to talk about GxE than it is to find it. Rather than trying to shoot the messenger because of his message, it would be far more useful to collect empirical data (not thought experiments about gravity) that demonstrate the importance of GxE. My reaction to Wahlsten's article is that when all is said and done, not much new is said and even less is done about identifying specific GxE.

Inherited quality control problems

Peter H. Schönemann

Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907

Electronic mail: *kc@brazil.psych.purdue.edu*

> Ironically, despite Mr. Justice Holmes' assertion in Buck v. Bell that "three generations of imbeciles are enough," there is some evidence to suggest that Carrie Buck's daughter Vivian, who was only one month old at the time she was appraised as "mentally defective" by a nurse, was in fact very bright
>
> (Areen 1985, p. 835).

Wahlsten tells us that a whole century of heritability estimates may have to be debunked because behavior geneticists overlooked the fact that interaction tests lack power. As he also notes, the power issue goes back at least to 1935. Thus he confronts us with yet another installment in a long series of interrelated revelations about mental tests which began in earnest when Kamin (1974) discovered that Burt's twin data were faked. It continued with Pike (1978) and Slack and Porter (1980), who reported that the SAT is much more coachable than we had been led to believe and Flynn (1987) who found that IQs, far from holding constant or even declining as we had repeatedly been warned, actually show massive gains. Crouse and Trusheim (1988) informed us recently that the incremental validity of the SAT over high school GPA is virtually zero.

The special twist of Wahlsten's story is that it implicates a basic statistical issue, which, moreover, had been discussed since 1935. How then can behavior geneticists still be confused about it? What is so difficult about appreciating the difference between a false positive and a false negative, say, of an AIDS test? As it turns out, behavior geneticists are not alone in their confusion about such elementary statistical issues:

> Questions of significance level versus power are indeed complex and not susceptible to easy answers. I shall circulate your letter back to the Associate Editor to see if he has comments on this aspect of your letter (Editor of the *Journal of the American Statistical Association*, personal communication, 1980).

This exchange was triggered by a brief Note (Schönemann 1981) which contested Geweke and Singleton's (1980) claim that the likelihood ratio test in (unrestricted) factor analysis "has considerable power even when the sample size is only 10" (p. 136). I showed that this stunning claim was misleading because, to arrive at it, the authors had used totally unrepresentative communalities (in the high 90s) in their simulation. For communalities in the range of the actually published empirical factor analyses (which average in the 60s, e.g., French 1951), the power of this LRT barely exceeds the alpha level for sample sizes of N = 100, let alone for N = 10.

When the editor instructed me to address the effect of the communalities on the *null* distribution (!), I asked him to explain to me the relevance of the null distribution for tests of fit which lack power. He never did answer this question.

Lack of statistical quality control stretching over decades is the rule rather than the exception in the mental test field.

For example, I know of no prominent statisticians who raised fundamental objections against Holzinger's "heritability index" h^2 (Newman et al. 1937), which has been used without interruption from the 1930s to the 1980s in the belief that it estimates the ratio of the genetic variance over the sum of genetic plus environmental variance. As was recently shown in (Schönemann 1989a), it cannot possibly estimate this ratio because Holzinger had made a mistake in deriving it. As a result, h^2 contains no environmental variance at all.

Similarly, when Vetta (1981) tried to lodge legitimate criticisms of faulty claims by Jinks and Fulker (1970) – who, among other things, purported to be able to estimate the number of genes involved in IQ with uncanny precision: >22–100 (p. 348) – the editor of the *Bulletin* refused to publish his critique (Hirsch 1981, p. 23). No statistician in this country noticed anything wrong with the Jinks & Fulker paper, which Eysenck (1973) had praised as "the cornerstone on which any future argument about heritability must be based." More recently, Schönemann 1989a) showed (a) that the Shields (1960) data consistently violate several conditions implied by the genetic model Jinks and Fulker fitted to them, and (b) that a purely environmental model fits them better by a factor of 2.

What all these cases have in common is that (a) they all involve precisely defined statistical problems which have unambiguous answers, (b) faulty statistical claims often remain uncorrected for decades because editors refuse to publish valid criticisms,[1] and (c) statisticians are no better at quality control than psychologists. The present target article provides yet another illustration of this theme:

Its focal point, the lack of power of interactions tests, is a statistical issue par excellence. Presumably, then, at least some of the six-odd reviewers were statisticians. Why did none of them notice that Wahlsten's discussion of the power of interaction tests, his gravity example, his references to expected values, to Cohen's tables, and to noncentrality parameters suffer from the technical defect that they all refer to the wrong type of analysis of variance model? The models needed to justify the heritability computations, if any, are not fixed effects but *random effects* ("variance component") models which, under the alternative, do not involve the noncentral F-distribution at all. Rather, they involve the *central* F-distribution (e.g., Scheffé 1959, p. 244; Snedecor & Cochran 1967, p. 280).

This technical blemish in no way diminishes the importance of the power issue Wahlsten raised which, in fact, extends far beyond behavior genetics. However, it underlines the need to take a hard look at the present peer review system (cf. Peters & Ceci 1982; Harnad 1989) as a mechanism for ensuring statistical quality control. Recent events have heightened public awareness that faulty and fraudulent research claims, if left unchallenged over long periods of time, can have serious consequences.

NOTE

Whatever the reader may have gathered from Shockley's (1987) misleading Continuing Commentary, there never was any room for doubt that Jensen's positive Spearman correlations remain artifacts even under the changed (2-sample) definition of "Spearman's hypothesis", since both Guttman (1986) and I had submitted detailed evidence to this effect in two independent papers to BBS, which were both rejected. For further details see (Guttmann 1986; and Schönemann 1989b).

Variation in means and in ends

Arie J. van Noordwijk

Zoologisches Institut, CH 4051 Basel, Switzerland

Electronic mail: *noordwijk@urz.unibas.ch*

> Noah said to the animals: "Go and multiply." "We can't," said the adders, "being adders, we can only add." Noah then said: "No problem," and gave them log tables.

It is sometimes forgotten that a variance is an average squared deviation from *a* mean. In analyzing a problem, one should therefore carefully choose one's means. In an experiment one can even choose the deviations of the group means from the grand mean, and thereby choose part of the variance. In the end, we make statements about our own reality. This is nicely illustrated by the gravitation example Wahlsten gives. A large part of the variance of the product of two variables is correctly ascribed to the factor varying fourfold, and less to the factor varying less than twofold.

I would want to argue that if these relative variations in mass

and distance are representative for our world, then for many practical purposes mass and distance can be treated as additive. In a situation where different models cannot be discriminated with reasonable effort, a choice between them can become very important if we want to extrapolate. When we require interpolation only, structurally incorrect models with a good fit to data can still be very helpful. The past two decades have given us good examples of both in the field of macro-economics.

Many people, if not most, think in terms of addition. My best example to illustrate this comes from a practical joke I was marginally involved in. Living on the fourth floor of a six-story student residence, we stuck an index card in the elevator with the text: "Button 6 out of order, press 2 & 4." Some victims used the stairs from the fourth floor up for days. It would be most interesting to turn this joke into an experiment and to compare the reactions to this obviously false additivity in departments of psychology, physics, biology, engineering and law.

I heartily concur with Wahlsten that (a) it is silly to conclude that two factors are "truly" additive if an interaction is not demonstrable and (b) that one should have a realistic idea of the power of tests. Problems of interpretation arise from both under- and overpowered tests. It is useful to point out the relatively low power of tests for interactions. What I miss, however, is general advice. Creating a greater diversity of circumstances, not only through extending the range of variables, is probably the best general recipe for increasing the discriminatory power of tests.

Wahlsten creates the impression that it is criminal to treat the real world as additive. I see a contradiction between his treatment of transformations that might improve additivity and his stated aim of better understanding developmental processes. In my experience, knowing which transformations should be applied contains a wealth of information in itself. At the same time, the very same process may be additive or multiplicative depending on the level of observation. At the population level, reproduction is best described as a multiplicative process; at the individual level we normally ask different questions, e.g., questions concerning the timing of events. In this context the reproductive process can be adequately described by: 1 + 1 = 3 or occasionally 4.

What is required to study mechanisms? One gets the strong impression that Wahlsten's target article has a generally valid methodological point, namely, that the test for an interaction in an ANOVA is not so powerful, but that most of its space is devoted to a topical rather local debate. Creating a dichotomy between investigating mechanisms and calculating heritabilities is rather odd to me, because the very reason why my colleagues and I calculate heritabilities is to gain an understanding of mechanisms of micro-evolutionary change. My own field is evolutionary biology, and more specifically genetic ecology, which can be described as the application of quantitative genetics to natural populations. Natural populations tend to live in heterogeneous environments and to show nonrandom associations of genotypes over environments either as a consequence of selection or as a consequence of habitat choice. Moreover, natural selection is virtually always a consequence of changes in environmental conditions. The same environmental conditions may totally alter the expression of genetic variation. It is therefore likely that average heritabilities multiplied by average selection pressures give a false indication of the realized response to selection. Much depends on how selection operates. In my eyes it is likely that for adapted organisms, most traits are almost selectively neutral most of the time in most places. In tackling such problems, apportioning the total variance to genetic and environmental components under different sets of environmental conditions is a very helpful summary statistic. The changes in this summary statistic can discriminate among potential mechanisms.

Thus, a heritability estimate becomes as much an ecological parameter as a genetic parameter. One can take this a step further and play around with a subdivision of the environmental variance. Analogous to the customary subdivision of the genetic variance V_g into a number of components, the additive genetic variance V_a, the dominance variance V_d, interaction (epistatis) variance V_i:

$$V_g = V_a + V_d + V_i \ldots \quad (1)$$

it is possible to subdivide the environmental variance, for example, into a component due to temperature, a component due to food abundance and the rest due to unexplained environmental variance. The reason why this isn't customary is that quantitative genetics has been developed and is mostly applied in the controlled laboratory and agricultural environment. If we use F for known environmental factors, we could write our example as:

$$P = G + F_t + F_f + E \quad (2)$$

at the individual level.

Where the contribution to the phenotype from temperature conditions depends on the genotype, we could formulate the reaction norm for temperature as:

$$F_t = f(G,T) \quad (3)$$

or, in words: The effect of temperature on the phenotype is a function of the genotype and temperature. A simple function might be a sum of two independent effects. Using a sum is helpful in that it allows us to maintain a distinction between the mean effect of an environmental factor on phenotype- and genotype-specific deviations.

Conclusion. Quantitative genetics is very much a top down approach, using extreme simplifications to concentrate on the most important quantitative aspects. Of all the simplifications made and of all the distortions of reality that are thereby introduced, assuming additivity often adds only relatively minor errors. If some people conclude in some cases that their example shows true/real additivity, one should fulminate much more against the true/real than against the additivity.

Apart from the fact that in my area predicting the response to selection is directly relevant, but rather difficult, studying the ecology (in its literal sense of relations to the environment) of subdivisions of the phenotypic variance is an important way to gain insight into the mechanisms of genotype-environment interactions. As far as I can judge, the latter would be equally valid in a psychological context. It is then probably most important to choose one's environmental means very carefully.

Recommended further reading: The Bioscience special issue on reaction norms (July/August 1989) with contributions by Stearns (1989), Scharloo (1989), Schlichting (1989), Dodson (1989), and van Noordwijk (1989) gives an overview of many different ways to study interactions of genes and environment.

Author's Response

Goals and methods: The study of development versus partitioning of variance

Douglas Wahlsten

Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

Electronic mail *userdlwa@ualtamts.bitnet*

The central thesis of the target article is that the statistical power of two-way analysis of variance to detect certain

kinds of interaction or nonadditivity is substantially less than the power to detect main effects in the same circumstances. Among the 26 commentators on this article, 24 express agreement in one form or another with this claim. This is gratifying, but not entirely surprising, because the point had been made some time ago by Neyman (1935) and has been reiterated from time to time in the literature. It is apparent that several commentators (**Dawes, Denenberg, Maxwell**) had already addressed this matter in their teaching and scholarly studies, if not in print. It must be admitted, however, that many researchers in the behavioral and brain sciences were not aware of the extent of the difficulty, or of its history. The target article demonstrates the magnitude of the difference in power specifically for a multiplicative model and several other realistic alternatives to additivity. This should help to create an awareness of the problem among readers, but, as **Lipp** predicts and several commentators confirm, it is not likely to dissuade the advocates of heritability analysis from practising their art. The various commentaries illustrate with great clarity how the different perspectives and goals of investigators condition their attitudes towards statistical methods.

1. Questions of power and light. Several commentators argue persuasively that the problem of low power could be studied or expressed differently and perhaps more simply.

Dawes demonstrates clearly how the linear contrast for interaction in a 2 × 2 design can help us compare different types of interaction and their main effects. Using this approach with the Y = (jh)(ke) model when h = e = 1, the means would be

	H_1	H_2
E_1	1	2
E_2	2	4

The contrast for the strain difference would be (4 + 2) − (2 + 1) = 3 and for H × E interaction would be (4 − 2) − (2 − 1) = 1. This humble example does get at the essence of the matter. I believe (with Lachenbruch 1988) the sample size required to achieve a specified degree of power should be inversely proportional to the square of the contrast, which implies that about nine times as many observations per cell will be needed to detect the multiplicative interaction compared to the main effect (see Wahlsten [unpublished]). The relative sample sizes indicate the additional quantity of subjects, time, and grant funds needed to detect real nonadditivity in the data. Dawes also presents valuable advice on the proper way to code interactions in multiple regression analysis.

Denenberg expresses the problem in terms of the correlation between group means across the various treatments. When two rat strains respond the same way to several environments, the correlation will be very high, and the mean square for interaction according to his equation (3) will approach zero. If there is no correlation, as when one strain is strongly affected by the treatment and the other shows no change at all, interaction should be substantial. This should apply to the scenario portrayed by **Kline**, in which one main effect is much larger than the other but the interaction is significant. If they respond in opposite ways, yielding a negative correlation, interaction should predominate. Denenberg suggests that the power of the test of interaction will be lower than for the main effects only when the correlation is positive. However, my example of Y = a + bX, Case 1, in Figure 2(a) of the target article suggests that the power to detect the interaction can be lower than for the main effect when the correlation is negative. Denenberg's formula (3) applies to sample data, whereas power calculations require population parameters derived from a model specified a priori. As suggested by **Maxwell**, the respective degrees of freedom are also part of the story; that is, the *number* of strains and treatment conditions as well as the correlation across treatments must be considered.

Maxwell casts further light on the subject by showing that the ordinal versus disordinal distinction between kinds of interactions does inform us about relative power in the specific instance of the 2 × 2 design when the degrees of freedom for main effects and interaction are identical. I agree with his contention that, for larger designs, the power of the test of interaction can be increased by using a planned contrast to test for an optimal interaction subeffect. Presuming that the test is planned before peeking at the results, this is feasible only when one has a good idea about the nature of the biological or psychological processes involved. Those aiming simply to partition variance may be stuck with the feeble global F test of interaction.

Cicchetti proposes that computer simulation be used to explore further aspects of the problem which may not be amenable to the technique I used. **Chiszar, & Gollin** and **Maxwell** make a similar suggestion. The Soper et al. (1988) and Adams et al. (1985) studies mentioned by Cicchetti as well as the study by Heth et al. (1989) provide excellent examples of the utility of this approach. When I first presented a paper on this topic at the Behavior Genetics Association meeting in Minneapolis (Wahlsten 1987a), I used a Monte Carlo program written in "C" to obtain a quick estimate of relative power for a 5 × 5 design. Subsequently, I adopted the algebraic method because of its greater apparent elegance, but I acknowledge that there will be trouble extending it to situations where a computer simulation would work readily. For example, Chiszar & Gollin point out that ANOVA main effects and Type I errors may be relatively robust against nonnormality or heterogeneity of variance, but that these issues have not been well evaluated with regard to interaction or type II errors. These and other violations of assumptions could be incorporated into a Monte Carlo study, with due attention to the properties of the random number generator (Press et al. 1988, Chap. 7).

Several commentators recommend alternative approaches to the standard ANOVA rather than increasing sample size. **Bullock** maintains that the larger n does not solve the root problem afflicting behavior genetics; he calls for the use of consistency checks on the results of ANOVA and better training of psychologists in applied mathematics. His remedies have considerable merit. I contend that if one wishes to rely on the results of ANOVA to assess interaction, larger samples ought to be used. At the same time, in section 12 of the target article I note that the global F-test is not necessarily the best

solution. **Carlier & Marchaland** propose Bayesian inference as a cure for some of the shortcomings of ANOVA, because decisions about significance are contingent upon effect size. Bayesian methods also incorporate explicit statements about the investigators' beliefs (Berger & Berry 1988) and thereby discourage impetuous acceptance of the null hypothesis. Both consistency checks and Bayesian inference can help to avoid the worst pitfalls of ANOVA, as can the likelihood ratio test advocated by Marler (1980) as well as by **McGuffin & Katz. Goodall** informs us of a forthcoming volume on ANOVA which addresses these issues directly, so that our minds can remain open to new procedures. However, I doubt that any mathematical method can obviate the need for larger samples when one is seeking to make a finer discrimination or to test for subtler effects such as multiplicative interaction. Better math can increase efficiency and reduce the drain on scarce research funds, but this cannot equate the sensitivities of even the best test to large and small effects.

Of course, extending the range of circumstances to achieve higher power, as suggested by **Henderson** and by **Van Noordwijk,** could preclude the need for larger samples if the experiment allows for this; but the power of ANOVA will still be lower for the interaction than the main effects if H and E are multiplicative.

2. Generality of the model. Two commentators (**Kempthorne, Schönemann**) assert that a random-effects rather than a fixed-effects model should have been used. Although neither claims definitively that the problem of low power to detect interaction would disappear with a random-effects model, they suggest that the central thesis is not established beyond doubt in the target article. Neither elaborates reasons why one model is preferable. Having reconsidered the question, I still think the fixed effects model is appropriate for the present purpose. Furthermore, the central conclusion would not be altered by using the random effects model; on the contrary, the power to detect heredity by environment (H × E) interaction would be even lower than with a fixed-effects model.

The main issue in choosing the model is the generality of the results. With fixed effects involving, for example, two inbred mouse strains reared in two environments, the results must be considered specific to the strains and environments actually studied. On the other hand, if several genotypes are sampled randomly from a larger population of diverse genotypes, as proposed by **Kempthorne,** and several rearing environments are similarly chosen, then results can legitimately be applied to the entire population.

Let us ask: How do the users of two-way factorial designs actually choose their animal subjects and levels of environment, and what sorts of generalizations do they make? As **Henderson** confirms, when rodents or flies are the subjects, it is customary and wise to choose strains likely to have extreme scores or, as advocated by Ward (1985), known to differ greatly at a large number of genetic loci. It is common practice to test a wide variety of strains and then choose two with extreme scores for further genetic analysis (Bauer & Sokolowski 1985; Wimer & Wimer 1982). Alternatively, selective breeding may be used to produce a maximum difference in behavior (Brush et al. 1985; Ricker & Hirsch 1988). Environments are typically chosen to yield a large difference in outcome, and care is exercised to restrict conclusions to the conditions actually observed.

The 2 × 2 design in particular inherently lacks generality. Even if the uninitiated were to choose strains and levels of environment entirely at random with the most inscrutable of computer programs, I cannot imagine even one experienced researcher accepting the results as representative of a wide range of strains and treatments.

When two- and three-way factorial designs are used, relatively few levels of each factor are commonplace. It is not at all surprising therefore that tabulations of power of ANOVA, such as those by Cohen (1988) as well as Rotton and **Schönemann** himself (1978), often restrict attention to fixed effects models.

If the design of the experiment does indeed warrant use of a random effects model, the power of the test of H × E interaction can be readily estimated. Koele (1982) considers a two-way design where the significance of the interaction term is tested against the error mean squares using the critical ratio F_c. If the true variance of interaction effects is σ^2_{AB} and the error variance is σ^2_ϵ, the result is

$$\text{Power} = \Pr\{F \geq F_c/(1 + n\sigma^2_{AB}/\sigma^2_\epsilon)\}.$$

In the target article I follow Cohen's convention for effect size f as the ratio of standard deviations. For two-way interaction $\sigma_{AB}/\sigma_\epsilon = f_I$. Thus,

$$\text{Power} = \Pr\{F \geq F_c/(1 + nf^2_I)\}.$$

Although **Kempthorne** suggests that the power of the interaction effect in a random effects design can be assessed with Tang's tables of the noncentral F distribution, Scheffé (1959, p. 227), Koele (1982) and **Schönemann** argue that the central *F* distribution is appropriate in this situation. Let us now compare the power of the same interaction effect size estimated (a) as in Cohen (1988) under the fixed effects model with the noncentral *F* distribution, as done in the target article, and (b) under the random effects model as in Koele (1982) with the central *F* distribution. Let there be five levels each of heredity and environment with 10 subjects in each of 25 groups, and let $\alpha = 0.05$. The power to detect H × E interaction is generally lower under the random effects model.

The test of main effects under a random effects model is properly done with respect to the interaction mean squares if the interaction is indeed significant. This outcome will have devastating consequences for the power of tests of main effects; but it will occur rarely because of low power and is not pertinent to the theme of the target article, where the focus is on situations in which there really is interaction but the researcher concludes that there is none. In such a situation, one may decide to test the main effects against the error term, a step to be taken with trepidation (Hays 1988) although it is often taken in practice. Suppose the main effect size for heredity or a strain difference is 0.4 in a 5 × 5 experiment with 10 mice per group. As shown in the target article, under a fixed effects model the power to detect the main effect will be greater than 99%, whereas with random effects it will be 46%. Multiplicative interaction with equally spaced levels of H and E will have a corresponding effect size of 0.189, which will yield power of 36% under a fixed effects

Table 1. *Power (%) of 5 × 5 ANOVA to detect interaction under fixed effects and random effects models with same effect sizes when n = 10 and α = 0.05.*

Interaction Effect Size	Fixed effects	Random effects
0.05	6	6
0.10	11	9
0.20	41	27
0.30	84	59
0.40	99	84

model and 24% under random effects. That is, under a random effects model the power to detect the H × E interaction is considerably lower than to detect the main effect of strain, but the difference in power is not as great as with fixed effects. The central thesis of the target article is supported under a random effects model as it is usually applied.

Because a fixed effects model is appropriate to the target article and conclusions are the same under a random effects model in any event, **Schönemann**'s allegation about lack of "quality control" on the part of the eight (not six) *BBS* referee reports is without merit. It strikes me as bizarre how Rotton and Schönemann (1978) once stated "in factorial designs interaction tests seldom match the power of main effect tests" and provided an illustrative example, yet now Schönemann neither claims priority nor repeats this unequivocal remark.

3. Complexity of the models. Several commentators suggest that a more complete model of heredity and environment should be used. If one were to propose a viable model of human behavior, without doubt several additional processes should be considered. The two-way factorial design with genetically uniform inbred strains randomly assigned to extreme environments as outlined in the target article is aimed expressly at avoiding certain complications in order to reveal the fundamental problem of insensitivity to interaction in the clearest possible way.

Crusio, Dawes, Dudley, Harrington, Kempthorne, Kline, and **McGuffin & Katz** argue that heredity and environment can be correlated, and that such covariance can be important in a realistic model of behavior. For the case of human society or animal populations outside the laboratory, I concur. Confounding of H and E, an extreme form of covariance, can make the estimation of some parameters impossible, and covariance occasioned by genetically influenced behavioral modification or choice of the environment can wreak havoc in a path model. As Goldberger (1978) and Taylor (1980) have shown, even minimally complete models suffer from underdetermination in which there are more parameters to be estimated than there are observed correlations available. Harrington also suggests that structural equations can yield biased results when two regression coefficients are opposite in sign, so that interaction will fail to appear regardless of sample size. The critique by Kempthorne (1978) of prevalent misconceptions is most informative, as is the review of additional criticisms by **Dudley** and **Hirsch.** Let me assure these astute commentators and *BBS* readers that I do not think H × E interaction poses the only challenge to heritability analysis. The two-way factorial design using inbred strains minimizes covariance, as **Crusio** notes; and omitting such an effect from the ANOVA model can be justified in the target article and in the laboratory when an experiment of this kind is done.

Kline questions the relevance of studies of nonhuman animals to humans because "man alone" has a strong influence on the environment. My reading of behavioral ecology suggests otherwise. **Bullock** states that humans are the "most extreme" in this respect but are not alone. The very essence of animal behavior is transformation and creation of the environment, as should be apparent in the cycle of ingestion, digestion, and excretion, as well as in the phenomena of habitat selection, burrowing or nest construction, etc. **Chiszar & Gollin** stress the "interdefinition of genome, organism, and ecosystem." Lewontin (1982) also explains very well the interpenetration of the organism and its environment. In the laboratory we attempt to restrict the operation of some of these processes in order to simplify and analyze mechanisms. Just as Mendel needed a uniform plot of ground to reveal laws of genetic transmission, so is precise control of heredity and environment in the lab helpful to document interaction. In society at large we should expect to find both Mendelian inheritance and heredity-environment interaction, and we should be skeptical of any model which presumes the presence of one but the absence of the other.

Dudley and **McGuffin & Katz** propose that a realistic model ought to incorporate interactions between genes at different loci (epistasis) as well as between genes and environment. This is especially important when one wishes to analyze the components of global heredity, which can only be done with cross-breeding schemes. Comparing several inbred strains varies heredity but cannot further elucidate its mechanisms. As mentioned by **Crow,** there is abundant evidence that the consequences of genes at one locus depend on genotype at other loci. This has been amply demonstrated for mouse pigmentation (Lamoreux & Pendergast 1987), obesity-diabetes (Coleman 1981), and brain development (Billings-Gagliardi & Wolf 1988; Kerner & Carson 1986), and evidence is sometimes found for behavior (Bateson & D'Udine 1986). When one is attempting to understand the dynamics of development, these phenomena can be most informative. Contrary to the claim by Crow that inbreeding reduces error variation, inbreeding often increases phenotypic variance above the level seen in F_1 hybrids (Hyde 1973; Palmer & Strobeck 1986), and this may very well stem from epistatic interaction.

If a more complex model were formulated to take account of several kinds of covariance and gene-gene interaction as well as heredity-environment interaction, the sensitivity of the test of interaction would very probably be extremely low – if a decisive test could be formulated at all for humans. The 2 × 2 test proposed by Plomin et al. (1977) is not valid because it classifies adoptees by phenotype of biological parents rather than the adoptees' own genotypes. Any test of G × E interaction must

involve replicated genotypes reared in different environments. Otherwise, innumerable combinations of genes and environments can yield the same phenotypic outcome. The target article concentrated on the 2 × 2 test of interaction because it is advocated by well-known spokesmen for behavior genetics. The point is made that even if one considers this test credible for humans (which I do not) the power is usually so low that any negative results must be inconclusive. The great confidence with which some researchers proclaim the absence or rarity of H × E interactions is not warranted by the shaky ladder of dubious assumptions on which they stand.

4. Focus on development. The target article contrasted two research agendas: the study of development and the partitioning of variance. The commentaries indicate that this dichotomy was no figment of my imagination. Developmentalists have very little interest or faith in assigning definite percentages of variance to contributing causes. Inferential statistics may be used as an aid to decision making, but the statistical models are not reified the way they so often are in behavior genetics. The developmental perspective in no way denies the importance of genes or espouses "genetic nihilism," as implied by **Henderson.** Interactionism advances beyond the sterile nature-nurture dichotomy inherent in the ubiquitous G + E of human behavior genetics.

Bookstein stresses that "the scientist needs to know the form of f," the function relating heredity and environment. **Bullock** says we need "quantitative models of actual processes." **Cheverud** wants us to improve the genetic analysis of development by generating "developmental models which would help guide the analysis." **Chiszar & Gollin** seek "the adduction of the organizing principles that mediate development." **Crusio** urges the use of genetic techniques for studying causal relationships between brain and behavior. **Harrington** instructs us that "for explanatory purposes a model must be logically and theoretically, not empirically, based," and he shows how a developmental perspective can specify the proper order of entry of variables into a multiple regression equation. **Lipp** argues that "understanding the target" of gene action should precede studies of how genes affect it. **Maxwell** stresses the importance of developing a "correct model of the phenomenon under study." **Nyborg** is after precise knowledge about "the character, mechanisms, and locus of action" of causes, and he is not satisfied with gene effects that are "assumed rather than localized and specified."

Explorations of physiologically real interactions can lead to noteworthy progress in our understanding of ontogeny (e.g. Beardsley 1988; Ingham 1988; Vardimon et al. 1988; Yeakley et al. 1987). The hypotheses proposed by **Lipp** and **Nyborg** exemplify the fruitfulness of a developmental perspective. Lipp proposes that later acting "psychogenes" will tend to affect neural systems at a higher hierarchical level, and that single gene effects are more likely to be observed on complex behaviors. This approach to behavior genetics is meaningful for the developmentalist. Nyborg envisions bidirectional relations in a way that informs rather than offends the psychologist interested in chemistry and behavior. Dynamic nonlinear models of gene action in neuroendocrinology illustrate the merits of a physiological and psychological interactionist approach (see Nyborg).

Crow draws a distinction between H × E interaction which can and cannot be removed by a monotonic transformation and regards only the latter as "real." However, to the developmentalist the judgment of which is real and which is mere appearance cannot be made simply from mathematical form. One must know how things actually work. Only then can certain functional relationships be regarded as trivial instances of apparent interaction, while others are seen as profound. Likewise, the merits or demerits of the idea mentioned by **Goodnight** and championed by Falconer (1981) – that gene-environment interaction can be subsumed under genetic correlation as the proportion of genes acting in common in two environments to determine two measures – must be decided by molecular biology, not by population genetics. I, for one, very much doubt the veracity of this hypothesis because the calculations assume additivity of genes and environment *within* the corresponding environments, while allowing qualitative differences *between* environments.

Plomin's contention that main effects and interactions are "independent" makes no sense from a developmental perspective (Lewontin 1974; Oyama 1985). It is also mathematically wrong for the case of a random effect model that ought to be applied to a human population. The two-way fixed effects ANOVA we use in laboratory studies will separate the total variance into four separate piles even when the true functional relationship is multiplicative. Main effects are defined as being additive, and interaction is defined as the departure from additivity. If there is clearly significant departure from additivity, the null hypothesis of additivity should be rejected outright. Then there is no additivity at all and main effects are mere artifacts of the algebra. If H and E are multiplicative, then 100% of the relationship is H·E and none of it is additive.

Plomin's claim that in the target article the finding of interaction had no effect on the main effects is mistaken. As explained in section 7, I arbitrarily fixed the effect size f of the largest main effect at 0.4 for every model so that relative power could be better perceived. I also showed in section 7 that, under the Y = H·E model the difference between strain means in a one-way design depends on the environment in which they are reared. If J strains with values for heredity H equally spaced by h units are all reared in E_1 = e then the standard deviation of strain means is

$$\sigma_H = \frac{he}{2}\sqrt{\frac{(J+1)(J-1)}{3}},$$

which includes the e term reflecting the rearing environment. If they are instead reared in K different environments, as in the target article,

$$\sigma_H = \frac{(K+1)}{2} \cdot \frac{he}{2}\sqrt{\frac{(J+1)(J-1)}{3}}$$

That is, multiplicative interaction increases the main effect of strain by a factor (K + 1)/2. If this be seen as independence, then new spectacles are in order.

The developmentalist seeking to understand functional

relations needs better measurement of the fundamental processes, not just better statistics. **Bookstein** argues persuasively for this and provides instructive examples in his own applications of tensor biometrics (Bookstein 1987). I agree that two-way ANOVA does not really provide a *measure* of interaction in the strict sense. Perhaps we should name the interaction mean squares an indicator. Now, can we ever measure H, E and H × E interaction in their own units? For environment, many obvious examples of degrees of temperature, mg/kg for a dietary component, hours spent reading, and so on, are familiar. Current conceptualization and attempts at measurement of environment in human behavior genetics leave much to be desired (Wachs 1983), but proper measures are possible. Heredity, especially the genetic aspect, poses greater difficulty because it is inherently categorical. For example, the "jimpy" (*jp*) gene in mice differs from the normal allele by a substitution of an adenine for a guanine nucleotide base in the DNA (Nave et al. 1987). The genetic variable strongly affects the abundance of myelin proteolipid protein (Gardinier & Macklin 1988), but a measure of the protein is not a measure of the gene itself, nor is the amount or spatial distribution of myelin in the nerve bundle a measure of the gene because numerous other genetic loci and interacting physiological processes combine to govern the result (Lemke 1988).

Given that genotype must be a categorical variable, it can be represented by a dummy variable in a multiple regression equation for suitably designed experiments; and this can provide an index of H × E interaction. For example, Bulman-Fleming and Wahlsten (1988) found that the adult brain weights (measured in milligrams) of inbred BALB/cWah2 mice declined linearly with litter size (measured in number of pups). However, the slope of the relationship was significantly steeper when the mice were derived from ovarian tissue grafted into an inbred BALB/c mother (5 mg smaller for each additional pup) than when grafted into an F_1 hybrid mother (3 mg/pup). Is the slope of the line in mg/pup a measure of the maternal environment, whereas the 2mg/pup difference in slope measures interaction? Neither is a measure. What we measure is brain size of mice from litters of size 2, brain size of mice from litters of size 3, and so on. A slope of 5mg/pup is then an inference drawn with the aid of linear regression across litters. Because the difference in brain size between different litter sizes must represent an effect of environment, it may be seen as a pure indicator of environment alone, but the significant interaction alerts us to the fact that the value of 5mg/pup is strain dependent, as well.

A closely related matter is raised by **Carlier & Marchaland,** who point out that inbred strains differ in maternal environment as well as genotype and that consequently a strain-by-environment interaction cannot discriminate between gene-environment interaction and interaction with maternal environment. They are absolutely right about this. **Henderson** fails to make this important distinction. I myself have emphasized the confounding of genetic (G) and E effects in strain studies (Wahlsten 1979), and results confirming the importance of maternal environment have been reported from my laboratory (Bulman-Fleming & Wahlsten 1988; Wahlsten 1983; Wainwright 1980). Is it fair and proper that I should now be hoist on my own petard? In the target article I was careful to talk about the heredity of a strain and to use the symbol H rather than the customary G found so pervasively in behavior genetics texts, but apparently I was not careful enough. There are passages, such as the second paragraph of Section 4, where the distinction between H and G is obscure. In a factorial experiment comparing inbred strains, heredity is "an operationally defined entity which includes the usual direct chromosomal influence . . . plus the less-widely recognized differences in maternal environment" (Wahlsten 1983, p. 220).

The studies by Carlier and her co-workers show how the maternal environment itself can interact with genotype. Psychologists are gradually becoming aware of the importance of the very early environment for behavior (e.g., Gutzke & Crews 1988; Smotherman & Robertson 1988). The comments of **Carlier & Marchaland** as well as **Harrington** pose a challenge for human adoption studies because adopted away twins always share a prenatal environment and the singleton adoptee spends at least nine months in an environment provided by its genetic mother. As if things were not already complicated enough, we must now be aware of maternal effect genes (Winslow et al. 1988) and chromosomal imprinting (Reik et al. 1987; Sapienza et al. 1987). *C'est la vie!*

Crow and **van Noordwijk** suggest that gene effects may be highly nonadditive at one level, the molecular level in particular, and yet appear additive at another level. Although these points are made in defense of heritability analysis, they warrant careful study on their own merits because of the developmental content. I cannot agree with Crow's strong statement that "tiny increments of *anything* are additive." This fundamental theorem of calculus cannot explain the behavior of certain deterministic but nonlinear dynamic systems of interactions under conditions far from equilibrium – the realm of so-called "chaos" (Gleick 1987; Grebogi et al. 1987). Such systems are characterized by a sensitive dependence on initial conditions that makes long-term prediction of macroscopic behavior such as weather all but impossible from the standpoint of a collection of locally acting causes.

Hyperion, a moon of Saturn, has an orbit so chaotic that a measurement accurate to 10 digits would not be sufficient to know its location, even crudely, a mere two years later (Killian 1989); it is safe to say that it will still be orbiting Saturn, but for many asteriods the limits of their orbits are not assured. The history of physical science on the earth might read quite differently if Newton had sat beneath a lurching, irregular Hyperion among a multitude of moons rather than the spherical solitaire he knew so well. I suspect that **Crow** and **Van Noordwijk** are right about definite nonadditivity at one level and apparent additivity at another, and that chaos theory will prove applicable to embryonic and social development alike. If so, any claim that valuable information about mechanisms of development can be gleaned from patterns of correlations among measures of the outcomes of development is dubious. [See also Skarda & Freeman: "How brains make chaos in order to make sense of the world" *BBS* 10(2) (1987)]

5. Should ANOVA be banned? Although **Bookstein** seems to advocate a ban on ANOVA, I do not. **Plomin** attributes to me the view that, because of the low power of ANOVA to detect interaction, the traditional analysis of variance model should be abandoned and the messenger shot. **Detterman** echoes this view.

My opinion, as stated in the target article, is that (a) if a researcher wants to use two-way ANOVA to test additivity, then a large sample size should be used to insure sufficient power for the test of interaction, and (b) if the functional relationship between two factors such as H and E is nonadditive and perhaps multiplicative, *then* it makes no sense to ascribe a definite percentage of the total variance to mutually exclusive and independent causes. Thus, I maintain that the ANOVA is appropriate in some situations but not in others. In section 13 of the target article, I suggest that it is useful in the "early phases of investigation." While sympathizing with the view of **Cheverud** that the main problem is not so much with ANOVA as with developmental theory, I am sensitive to the criticism by **Bullock** that calling for larger sample sizes may serve to perpetuate the "undue hegemony of ANOVA" in psychology.

Especially in a one-way research design, partitioning the variance can be very helpful. Knowledge that cognitive gender differences sometimes account for a paltry 1% of total test score variance (Hyde 1981) and that the magnitude of the difference is approaching the vanishing point (Feingold 1988) can and should inform current debate about gender discrimination in education and hiring. Estimating the strength of an effect, which implicitly requires a partition of variance, allows us to make a wise choice of sample size on the basis of power calculations. When two or more factors vary simultaneously, however, the difficulties with ANOVA multiply and the tidy division of variance loses credibility.

6. What good is heritability? While advocating a flexible approach to data analysis and ANOVA, I, along with **Hirsch,** see the net contribution of the concept of heritability in behavior genetics as negative. **Bookstein** also expresses concern that without a better understanding of functional relations path modelling "is actively misleading." **Kempthorne** urges that most of the literature on heritability of human behavior be ignored. Other commentators, however, offer a spirited defence of this controversial h^2.

Heritability is central to human behavior genetics as it is commonly practised, but **Bullock** finds it difficult to believe my representation of the field is accurate. I direct him to the comments of **Crow, Detterman, Henderson** and **Plomin,** and to the leading text by Plomin et al. (1980), for up-to-date examples. **McGuffin & Katz** wonder whether any behavior geneticist really sees the estimation of h^2 as an end in itself or cares whether the value is 0.8 or 0.6 or 0.4. Here we can examine the two most recent semi-official overviews of the field in the *Annual Review of Psychology.* Henderson (1982) presents numerous estimates of the heritability of IQ, personality, and psychopathology. He gives a tentative estimate that narrow heritability of IQ might be "between .3 and .6," and he points out that more recent studies using better methodology yield lower heritability values. Loehlin et al. (1988) question Henderson's interpretation of history and, on the basis of even more recent data, assert: "It now appears that heritability estimates of general intelligence are back up again . . ." (p. 103) and have "reversed the trend toward lower heritability estimates" (p. 104). Peer commentary on the Plomin and Daniels target article in *BBS* [10 (1) 1987] also reveals some who care a great deal about the precise numerical value.

It is sometimes suggested that heritability analysis is justified because a finding of H × E interaction is a rare occurrence. **Plomin,** echoed by **Detterman,** argues this point. **Carlier & Marchaland** note the low prevalence of reports of genuine gene-environment interaction in behavior genetics. In weighing the evidence, we should keep in mind that, in the pages of a journal like *Behavior Genetics,* few serious attempts to test for H × E or GxE interaction are to be found at all. The most common design used by laboratory researchers in this field is a genetic crossing or selection experiment with all subjects reared in similar circumstances. We must also contend with the zeitgeist in the field so well satirized by Salsburg (1985); "having a significant interaction is a little like eating chicken with your fingers in public or wearing track shoes to a wedding. Somehow it is all your fault, and you are not quite sure what you have done wrong." He claims that "editors scream your experiment is no good" when presented with interactions. I have heard tales of this happening to my friends, and it has certainly happened to me. There may be a reporting bias in the literature.

Among active behavior genetics researchers, our two most senior colleagues, Benson Ginsburg and John Paul Scott, began their studies on mouse social behavior with discrepant findings (Ginsburg & Allee 1942; Scott 1942) produced by interaction with rearing and testing conditions (Ginsburg 1967). Ginsburg has been an interactionist even since, so we might posit a critical period for acquiring a developmental perspective. In any instructive review of research on early experience and mouse strains, Erlenmeyer-Kimling (1972) noted that "gene-environment interactions are numerous and . . . treatment effects are frequently reversed in direction for different genotypes." (p. 201) I have not done a rigorous count, but I am impressed by the large number of my colleagues who have reported or discussed interactions recently (e.g. Crabbe et al. 1988; Donovick & Burright 1984; Goodlett et al. 1987; Graf 1987; Satinder & Sterling 1983; Wilson & Sinha 1985; Zacharko et al. 1987). Others are mentioned in the target article. **Detterman**'s claim about "the absence of persuasive data indicating that interactions are important to a behavior genetic model" is an incredible statement that is a denial of the literature, not a review of it. Those who are blind to the existence of interaction will never progress to a discussion of more challenging issues, such as the distinction between interaction and separability mentioned by **Bookstein** and elaborated by Gregorious and Namkoong (1987).

In future, anyone who does review the literature on H × E interaction would be well advised to assess not only whether the interaction F ratio was significant at $\alpha = 0.05$ but what the power of the test was. Examples in the target article suggest that reports of significant interaction will tend to be infrequent even when H and E are not additive. When power is low, results will also be difficult

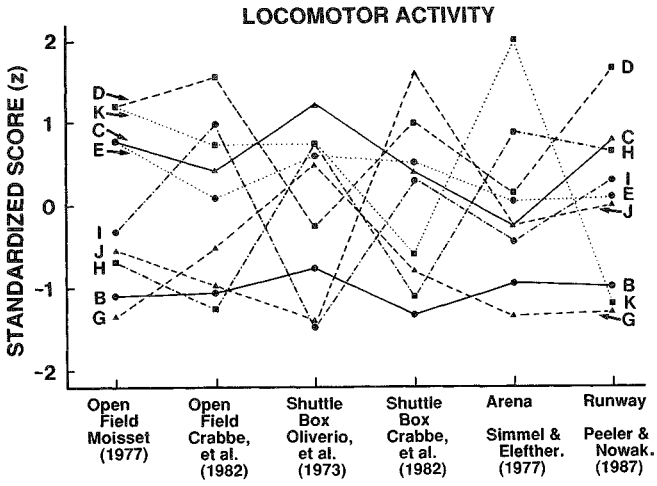


Figure 1. (Wahlsten) Locomotor activity in five studies of the Bailey recombinant inbred mouse strains and their two progenitor strains. In each study, the standard deviation of the nine strain means was determined. The value of *z* represents the number of standard deviations by which a strain mean differed from the mean of all nine strains in the study in question. The various studies assessed locomotor activity in either an open field (Crabbe et al. 1982; Moisset 1977), a shuttle box without electric shock (Crabbe et al. 1982; Oliverio et al. 1973), an arena (Simmel & Eleftheriou 1977) or a straight runway (Peeler & Nowakowski 1987). Abbreviations: B, BALB/cByJ; C, C57BL/6ByJ; D, CXBD/By; E, CXBE/By; G, CXBG/By; H, CXBH/By; I, CXBI/By; J, CXBJ/By; K, CXBK/By.

to replicate, so **Plomin**'s claim that failure to replicate plagues the H × E literature supports my thesis. Even more telling is a point he failed to mention: Main effects often do not replicate either. That is, the rank orders of strains given similar tests of behavior often differ greatly between laboratories (Wahlsten 1978). Figure 1 presents standardized mean values of motor activity of the seven Bailey (By) recombinant inbred strains and their two progenitors in five laboratories. The C57BL/6ByJ strain is generally high and BALB/cByJ is generally low, but the recombinants show far less consistency. Such discrepant results indicate a strong interaction of strain with either rearing or testing conditions. Peeler (1986) has shown that in the same laboratory the apparent genetic influence on activity depends on the time of day when testing is done and prior experience in the apparatus. As **Hirsch** points out, only a small fraction of possible interactions can be seriously assessed in our experiments, but those not scrutinized in a factorial design can intrude nonetheless and make replication in different situations quite unreliable. **Kline** contends that low power to detect interaction does not mean it must be present everywhere. None will disagree. However, low power does warn us that failures to detect or replicate an interaction do not prove that the factors are additive.

A large number of informative interactions has been reported with impressive consistency in the field of teratology as well as in studies of mouse genes such as "viable yellow" (Frigeri et al. 1988), "obese" (Bellward and Dauncey 1988) and "staggerer" (Guastavino 1988), not to mention strains of wheat (Roberts & Larson 1985) and Mendel's favorite, the garden pea (Reid & Murfet 1977). Numerous studies of developmental genetics support **Bullock**'s argument that H × E interactions are *likely* to occur in nature. **Goodall** also says we should expect to find some interaction.

Given this, we could insist that the statistical hypothesis of nonadditivity be treated as the null hypothesis and that additivity be seriously entertained only when the null hypothesis can be firmly rejected. This would not require any revolution in mathematics. Simply do a transformation of the data before *every* ANOVA and ask whether the interaction is significant. If H and E really are additive, this should create an interaction that could be removed by a suitable reverse transformation. On the other hand, if power were very low, the transformation would be of little consequence either way. **Goodall** proposes that transformation be regarded as an integral part of data analysis and that alternative models with and without transformation be *routinely* considered in studies of heredity and environment. This could strengthen many studies. My opposition to ad hoc transformations applies to studies where only one analysis of a transformed measure is presented to the reader as the only reasonable interpretation of the data.

Detterman, Henderson, and **Plomin** assert strongly that no H × E interaction has been demonstrated with humans. The immense literature on well-established Mendelian disorders refutes them decisively. For phenylketonuria and similar metabolic disorders, the child with two recessive genes is less able to regulate the levels of important substances in the blood and is *more sensitive* to variations in the diet. Whether the genetic defect would also interact with psychological treatments, as suggested by **Henderson,** is difficult to know because ethics forbid that a healthy child should be deliberately subjected to a poor education in the name of science. In the ethereal realm of "polygenic" behaviors where genes cannot be identified or counted, no conclusive test of specific GxE interaction exists. Nevertheless, in the study of psychopathology, it is often claimed that genotype determines susceptibility or vulnerability to the induction of psychosis by adverse experience (Kendler & Eaves 1986; Schulsinger et al. 1987; Tienari et al. 1987). If all is additive, why do psychiatrists bother with this blatantly interactionist theory?

When the possibility of H × E interaction is acknowledged, **Detterman** and **Plomin** insist that h^2 is still good because the interaction accounts for relatively little variance. **Crow** proves h^2 expected with additivity is only slightly less when a multiplicative model is used, and with several types of transformation. **Cheverud**'s position that interaction does not greatly bias the value of h^2 appears confirmed here, although there are other situations where the bias can be larger (Lathrope et al. 1984). Better support for the central thesis of the target article can hardly be imagined. Crow proves that the h^2 coefficient derived from parent-offspring regression is insensitive to the underlying structure of the data. Hence, h^2 cannot help us discover that structure or gain deeper insights into the nature of development. Looking at his model from the standpoint of an interactionist, let us inquire about the effects of changing the environment by the same amount for every child in the population. The additive model says that every child's phenotype should increase by the same amount. For the multiplicative

model, we need some numbers. Suppose both H and E factors have means of 3.0 and standard deviations of 1.0 in the population, and that H and E are uncorrelated. Now add 2.0 to each child's E value. For a child with a low H = 1.0, the phenotype will increase by only 2 units, whereas a lucky one with H = 5.0 will increase by 10 units, fully five times the increase for the less fortunate peer. Being oblivious to the functional relation between H and E, heritability analysis cannot predict the effect of changing the environment.

However **Crow, Goodnight,** and **Kempthorne** assure us that h^2 is useful in predicting the response to selective breeding, even when H and E are not at all additive. Sometimes the prediction is quite good, yet there are instances, especially with reproductive traits, when response to selection falls short of expectations from the h^2 value (Nordskog 1977). Wright's (1978) assessment of several decades of work on animal breeding also warrants caution. One particular shortcoming of the straightforward heritability approach is that it may hold for the first few generations of selection, but it cannot predict when or where the selection limit will be reached or whether there will be asymmetry of response in high and low lines. Strictly speaking, h^2 from parent-offspring regression really does not predict anything. Suppose we estimate h^2 by computing the regression of the mean score of the parents on the mean score of their offspring. Hill (1970) has demonstrated that the most efficient way to estimate h^2 is in fact to do a selection experiment for only one generation. Selection is essentially a parent-offspring regression study where the middle-scoring parents are not included. Perhaps one generation of selection can predict the response to selection in the next generation. This tells us simply that the properties of the population do not change much in one generation, whatever the true developmental relationship between H and E.

Cheverud and **Lipp** argue that calculation of h^2 can be a useful starting point because it provides evidence of genetic variation, whereas **McGuffin** and **Katz** recommend its use as a diagnostic aid in seeking forms of a heterogeneous disorder that may be heritable. Two objections come to mind. First, in studies of selective breeding or inbred strains, genuine chromosomal genetic effects cannot be distinguished from maternal environment effects (**Carrier & Marchaland**) or cytoplasmic inheritance. Likewise, twin studies are contaminated by covariance, cytoplasmic and uterine environment effects, and adoption studies are afflicted with covariance and stubborn maternal environment effects. Because h^2 is supposed to reflect only Mendelian gene effects in the numerator (Falconer 1981), it should not be reported unless other hereditary factors can be positively excluded. Second, evidence of putative genetic variation requiring further study can be shown with general purpose statistics. The approximate extent of hereditary variation among inbred strains or selected lines can be neatly summarized by estimated ω^2 (Hays 1988). If MZ twins show a much higher intra-class correlation than DZ twins or if parent-offspring regression is substantial, the presence of hereditary variation is a reasonable bet. Precisely what kind of hereditary mechanism are involved generally cannot be known solely from a strain or twin study. These kinds of experiments ought to be regarded as preludes to a more comprehensive study. Unfortunately, the h^2 coefficient is bound up with a very specific genetic theory, and citing a number for h^2 implies to many readers in psychology that they have just beheld the finale. What h^2 means to Falconer (1981) or Plomin et al. (1989) is automatically conveyed to a psychology student familiar with these authorities, even though the writer himself does not take the precise value of h^2 too seriously. What h^2 means to those with less education, I shudder to think.

Crow regards h^2 analysis as a useful, albeit indirect, approach to assessing the impact of environmental variation, e^2. In this vein, Heath et al. (1985) use h^2 values to draw conclusions about changes in schooling in Norway, and Plomin and Daniels (1987) cite results of path analysis to support their proposal that personality is unaffected by experiences common to members of a family. I cannot see how global statements about environmental variance tell us anything more about the specific actions of experience than the global h^2 tells us about the number, location, and physiological characteristics of relevant genes. As Wachs (1983) and **Bookstein** aver, to learn how environment affects the development of behavior we must have accurate measures of relevant features of experience. Furthermore, heredity and environment are not the only sources of individual differences in a population. Durable and noteworthy variations in the structure of an organism can emerge via processes internal to the embryo that are neither hereditary nor responses to local variations in the environment. (Kurnit et al. 1987; Lewontin 1982; Wahlsten 1987b; 1989b). With three sources of individual differences the potency of one cannot be specified by studying another.

Goodnight suggests that the absolute value of h^2 is not particularly useful but that the relative magnitudes of additive genetic and dominance variance have some relevance. **Crow** and **van Noordwijk** see a role for h^2 in evolutionary theory, and Crow thinks evolution must proceed gradually, in small, additive increments. On the other hand, several theorists see a close link between development and evolution, and, from this perspective, question the neoDarwinian dogma (Ho & Fox 1988). As suggested by van Noordwijk, phenotypic plasticity can play a very important role in molding an organism to its niche (Cavalli-Sforza 1974; Greene 1989); hence, gene-environment interaction should be central to evolutionary theory (e.g., Via & Lande 1985).

Finally, **McGuffin** and **Katz** suggest that my discussion of a link between heritability and eugenics sets up a straw man, and they claim no reputable human behavior geneticist would use heritability for eugenic purposes. It seems to me that the link between h^2 and selective breeding is inherent in a quantitative genetic model and has little to do with reputation. Because of this link, the heritability coefficient is not ethically neutral when it is computed for human IQ. Even if the scientists doing the computation are not proponents of eugenics, others may and probably will vulgarize their writings for eugenic purposes (Stein 1988). Within recent memory, some quite reputable geneticists *have* promoted selective sterilization of people with low intelligence (see Hirsch 1981; Nanney 1986). Today the Pioneer Fund seeks to advance its program of "racial betterment" (Lichtenstein 1977; May 1960) through large grants to several members of the Behavior Genetics Association (McCann & Currie 1989).

The stuff of controversy should not be brushed aside as mere straw.

7. Charges of bias. The target article, **Plomin** says, implies that behavioral geneticists, including himself "have not considered GxE" and have been "ignoring GxE." However, the target article in section 4 credits none other than Plomin et al. (1977) with proposing a formal 2 × 2 test of interaction. It cites five other items with him as senior author. My article states that in behavior genetics "the problem is not a lack of understanding about the importance of interaction in theory. Rather, there is a divergence of opinion about its occurrence in reality." The commentaries prove the latter point. As I also stated in the target article, the problem of low power is "generally understood by expert statisticians." A paper by Eaves et al. (1977) is cited as an example. **Henderson** quotes a sentence written by John Fuller in response to my 1979 paper to show how behavior geneticists supposedly knew all about interaction 10 years ago. However, my 1979 paper did not raise the issue of power of ANOVA to detect interaction. Furthermore, it is timely now to repeat Fuller's *next sentence* in that response concerning interactions: "It is good to be reminded that overlooking their existence may lead to faulty conclusions and premature generalization" (Fuller 1979).

Plomin states that I set out to denigrate heritability. It may appear that way to some people, so let me emphasize that the points made about the low power to detect H × E interaction apply to any kind of two-way or higher order interaction. **Hirsch** is right that assigning a percentage to environmental variance confronts the same problems as heritability. As **van Noordwijk** instructs, the same issues apply if we want to partition environmental sources of variance. **Goodall** stresses that ANOVA is perfectly general in this respect. **Dudley** presents three cells of an ideal adoption design with four groups. The fourth cell, children born into favorable homes but adopted into relatively poor homes, has now been filled in an excellent study done by France (Capron & Duyme 1989). The increase in IQ provided by going from a poor background to a favorable family environment was comparable to the decrease resulting from transfer in the opposite direction. The two-way interaction was not significant, but it suffered from the very problem of low power discussed in the target article.

8. Two kinds of gravity. The parody of Newton's law attracted some attention. **Lipp** was struck by the meaning that evidently missed **Detterman.** We should thank **Detterman** for his lovely graphs, because they bolster the conclusions I drew from Figure 2(a) in the target article. **Henderson** and Detterman want to see results when much wider ranges of mass and distance are used, so I provided the intrepid experimenter in section 6 of the target article with several burly assistants and sent them outside to a football field to gather data. Unfortunately, wind increased the error variance, and nothing was significant at all! **Kline** says we only need to know the mass of the relevant planet in order to calculate a person's weight because body size is constant throughout the universe. Given: I live in Edmonton, Alberta, on the planet earth. Can Kline therefore tell me my weight? **van Noordwijk** says a choice between models is determined by practical purposes and becomes most important when we want to extrapolate. Newton's law speaks to this matter, too. It could not be decisively verified by reference to the facts used to formulate the law, no matter how closely data and theory matched. It had to predict something new. This took place in 1846 when Leverrier used Newton's law to predict the location of an unknown planet from perturbations in the orbit of Uranus and then the existence of Neptune was confirmed by Galle and d'Arrest at the Berlin observatory using Leverrier's coordinates (Grosser 1979). What could be less practical than the exuberant delight of star gazers at this brilliant feat? It didn't really make much practical difference until the era of rocket journeys over 100 years later. Fortunately for us, the likes of Newton, Edmund Halley, and their followers would settle for nothing less than truth. On this day, August 25, 1989, the Voyager II space satellite flew past Neptune, guided by knowledge of a law of nature that predicted because it explained.

ACKNOWLEDGMENT

Skillful typing of this manuscript was done by Jan Zielinski.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.

Adams, A. & Bullock, D. (1986) Apprenticeship in word use: Social convergence processes in learning categorically related nouns. In: *The development of word meaning.* ed. S. A. Kuczaj & M. D. Barrett. Springer-Verlag. [DB]

Adams, K. M., Brown, G. G. & Grant, I. (1985) Analysis of covariance as a remedy for demographic mismatch of research subject groups: Some sobering simulations. *Journal of Clinical and Experimental Neuropsychology* 7:445–62. [DVC]

Alberch, P. (1983) Mapping genes to phenotypes, or the rules that generate form. *Evolution* 37:861–63. [DAC]

American Psychiatric Association (1980) Diagnostic and statistical manual of mental disorders, 3d ed. American Psychiatric Association. [PM]

Anastasi, A. (1958) Heredity, environment, and the question "How?" *Psychological Review* 65:197–208. [JH]

Areen, J. (1985) *Case material on family law.* Foundation Press. [PHS]

Atkinson, B. G. & Walden, D. B., eds. (1985) *Changes in eukaryotic gene expression in response to environmental stress.* Academic Press. [aDW]

Baker, B. O., Hardyck, C. D. & Petrinovich, L. F. (1966) Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens's proscriptions on statistics. *Educational and Psychological Measurements* 26:291–309. [DAC]

Bates, D. M. & Watts, D. G. (1988) *Nonlinear regression analysis and its applications.* Wiley. [CG]

Bateson, P. (1987) Biological approaches to the study of behavioral development. *International Journal of Behavioral Development* 10:1–22. [aDW, PM]

Bateson, P. & D'Udine, B. (1986) Exploration in two inbred strains of mice and their hybrids: Additive and interactive models of gene expression. *Animal Behaviour* 34:1026–32. [rDW]

Bauer, S. J. & Sokolowski, M. B. (1985) A genetic analysis of path length and pupation height in a natural population of *Drosophila melanogaster.* *Canadian Journal of Genetics and Cytology* 27:334–40. [rDW]

Beardsley, T. (1988) Developmental dialectics. *Scientific American* 259(Nov.):40–41. [rDW]

Bebbington, P. E., Brugha, T., MacCarthy, B., Potter, J., Sturt, E., Wykes, T., Katz, R. & McGuffin, P. (1988) The Camberwell collaborative depression study. 1. Depressed probands: Adversity and the form of depression. *British Journal of Psychiatry* 152:754–65. [PM]

Bebbington, P. E., Hurry, J., Tennant, C., Sturat, E. & Wing, J. K. (1981) Epidemiology of mental disorders in Camberwell. *Psychological Medicine* 11:561–79. [PM]

Bellward, K. & Dauncey, M. J. (1988) Behavioural energy regulation in lean

and genetically obese (*ob/ob*) mice. *Physiology and Behavior* 42:433–38. [rDW]

Benkel, B. F. & Hickey, D. A. (1987) A *Drosophila* gene is subject to glucose repression. *Proceedings of the National Academy of Sciences USA* 84:1337–39. [aDW]

Berger, J. O. & Berry, D. A. (1988) Statistical analysis and the illusion of objectivity. *American Scientist* 76:159–65. [rDW]

Bignami, G. & Bovet, D. (1965) Experience de sélection par rapport à une réaction conditionnée d'évitement chez le rat. *Comptes Rendus de l'Académie de Science* 260:1239–44. [H-PL]

Billings-Gagliardi, S. & Wolf, M. K. (1988) Shiverer*jimpy double mutant mice. IV. Five combinations of allelic mutations produce three morphological phenotypes. *Brain Research* 455:271–82. [rDW]

Blau, H. M., Pavlath, G. K., Hardeman, E. C., Chiu, C-P., Silberstein, L., Webster, S. G., Miller, S. C. & Webster, C. (1985) Plasticity of the differential state. *Science* 230:758–66. [aDW]

Bolles, R. C. (1988) Why you should avoid statistics. *Biological Psychiatry* 23:79–85. [aDW]

Boneau, C. A. (1960) The effect of violations of assumptions underlying the T test. *Psychological Bulletin* 57:49–64. [DAC]

Bookstein, F. L. (1987) Describing a craniofacial anomaly: Finite elements and the biometrics of landmark locations. *American Journal of Physical Anthropology* 74:495–509. [rDW]

Bouchard, T. J. & McGue, M. (1981) Familial studies of intelligence: A review. *Science* 212:1055–58. Tabulation sheets and lists of "Papers included" and "Papers excluded" (unpublished). [RMD]

Bovet, D., Bovet-Nitti, F. & Oliverio, A. (1969) Genetic aspects of memory and learning in mice. *Science* 163:139–49. [H-PL]

Bowers, K. E. (1973) Situationism in psychology: A critique. *Psychological Review* 80:307–36. [aDW]

Box, G. E. P. (1953) Non-normality and tests on variance. *Biometrika* 40:318–35. [[DAC]

(1954) Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* 25:290–302. [DAC]

Box, G. E. P. & Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society.* Series B 26:211–43. [aDW]

Box, G. E. P., Hunter, W. G. & Hunter, J. S. (1978) *Statistics for Experimenters.* John Wiley. [CG]

Broadhurst, P. L., Fulker, D. W. & Wilcock, J. (1974) Behavioral genetics. *Annual Review of Psychology* 25:389–415. [JH]

Brockington, I. F., Kendell, R. E. & Leff, J. P. (1978) Definitions of schizophrenia: Concordance and prediction of outcome. *Psychological Medicine* 8:387–98. [PM]

Brush, F. R., Baron, S., Froehlich, J. C., Ison, J. R., Pellegrino, L. J., Phillips, Sakellaris, P. C. & Williams, V. N. (1985) Genetic differences in avoidance learning by *Rattus norvegicus:* Escape/avoidance responding, sensitivity to electric shock, discrimination learning, and open-field behavior. *Journal of Comparative Psychology* 99:60–73. [rDW]

Bullock, D. (1987) Socializing the theory of intellectual development. In: *Meaning and the growth of understanding,* ed. M. Chapman & R. A. Dixon. Springer-Verlag. [DB]

Bulman-Fleming, B. & Wahlsten, D. (1988) Effects of a hybrid maternal environment on brain growth and corpus collosum defects of inbred BALB/c mice: A study using ovarian grafting. *Experimental Neurology* 99:636–46. [rDW]

Byerley, W., Mellon, C., O'Connell, P., Laloul, J.-M., Nakamura, Y., Leppert, M. & White, R. (1989) Mapping genes for manic-depression and schizophrenia with DNA markers. TINS 12:46–48. [H-PL]

Capron, C. & Duyme, M. (1989). Assessment of effects of socio-economic status on IQ in a full cross-fostering study. *Nature* 340:552–54. [rDW]

Carlier, M. & Nosten, M. (1987) Interaction between genotype and pre- or postnatal maternal environments: Examples from behaviors observed in inbred strains of mice. In: *Functional teratogenesis functional effects on the offspring after parental drug exposure,* ed. T. Fujii & P. M. Adams. Teikyo University Press. [aDW, MC]

Carlier, M., Nosten-Bertrand, M. & Michard-Vahnee, C. (in press) The separation of genetic from maternal effects. In: *Techniques for the genetic analysis of brain and behavior: Focus on the mouse,* ed. D. Goldowitz, D. Wahlsten & R. Wimer. Elsevier. [MC]

Carlier, M. & Roubertoux, P. L. (1986) Le développement des comportements: Effet des interactions entre le génotype et l'environment maternel. *Confrontations Psychiatriques* 27:63–88. [MC]

Carlier, M., Roubertoux, P. L. & Cohen-Salmon, C. (1983) Early development in mice I. Genotype and postnatal maternal effects. *Physiology and Behavior* 30:837–44. [MC]

Carroll, R. J. & Ruppert, D. (1988) *Transformation and weighting in regression.* Chapman & Hall. [CG]

Carroll, S. B., Winslow, G. M., Schupbach, T. & Scott, M. P. (1986) Maternal control of *Drosophila* segmentation gene expression. *Nature* 323:278–80. [aDW]

Carter, C. O. (1977) Letter. *Nature* 266:279. [RMD]

Castro, C. A. & Rudy, J. W. (1989) Early-life malnutrition impairs the performance of both young and adult rats on visual discrimination learning tasks. *Developmental Psychobiology* 22:15–28. [DAC]

Cavalli-Sforza, L. L. (1974) The role of plasticity in biological and cultural evolution. *Annals of the New York Academy of Sciences* 231:43–59. [rDW]

Cavalli-Sforza, L. L. & Feldman, M. W. (1973) Cultural versus biological inheritance: Phenotypic transmission from parents to children (A theory of the effect of parental phenotypes on children's phenotypes). *American Journal of Human Genetics* 25:618–37. [aDW]

Chauvin, R. (1977) *Ethology: The biological study of animal behavior,* translation. International Universities Press. [DAC]

Cheverud, J. (1984) Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology* 110:155–72. [JMC]

Cheverud, J. (1988). The evolution of genetic correlation and developmental constraints. In: *Population genetics and evolution,* ed. G. de Jong. Springer-Verlag. [JMC]

Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 65:145–53. [SEM]

(1977) *Statistical power analysis for the behavioral sciences,* revised edition. Academic Press. [aDW, SEM]

(1978) *Statistical power analysis for the behavioral sciences,* 2d ed. Erlbaum. [rDW]

Cohen, J. & Cohen, P. (1975) *Applied multiple regression/correlation analysis for the behavioral sciences.* Wiley. [RMD]

(1983) Applied multiple regression: Correlation analysis for the behavioral sciences, 2d. ed. Erlbaum. [RP]

Cole, W. A. & Trasler, D. G. (1980) Gene-teratogen interaction in insulin-induced mouse exencephaly. *Teratology* 22:125–39. [aDW]

Coleman, D. L. (1981) Inherited obesity-diabetes syndromes in the mouse. In: *Mammalian Genetics and Cancer,* ed. E. S. Russell & E. Schull. Alan R. Liss. [rDW]

Collins, R L. (1979) Selective breeding for the degree of functional lateralization in mice. *Behavior Genetics* 9:443–44. [H-PL]

Cooper, R. M. & Zubek, J. P. (1958) Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Canadian Journal of Psychology* 12:159–64. [aDW]

Crabbe, J. C., Deutsch, C. M., Tam, B. R. & Young, E. R. (1988) Environmental variables differentially affect ethanol-stimulated activity in selectively bred mouse lines. *Psychopharmacology* 95:103–08. [rDW]

Crabbe, J. C., Rigter, H. & Kerbusch, S. (1982) Analysis of behavioural responses to an ACTH analog in CXB/By recombinant inbred mice. *Behavioural Brain Research* 4:289–314. [rDW]

Cronbach, L. J. & Snow, R. E. (1975) *Aptitudes and instructional methods: A handbook for research on interactions.* Irvington. [DKD]

(1977) *Aptitudes and instructional methods: A handbook for research on interactions.* Irvington. [SEM]

Crouse, J. & Trusheim, D. (1988) *The case against the SAT.* University of Chicago Press. [PHS]

Crusio, W. E. (in press) Quantitative genetics. In: *Techniques for the genetic analysis of brain and behavior: Focus on the mouse,* ed. D. Goldowitz, D. Wahlsten & R. E. Wimer. Elsevier. [aDW, WEC]

Crusio, W. E., Schwegler, H. & van Abeelen, J. H. F. (1989) Behavioural responses to novelty and structural variation of the hippocampus in mice. II. Multivariate genetic analysis. *Behavioural Brain Research* 32:81–88. [WEC]

Darlington, R. B. (1968) Multiple regression in psychological research and practice. *Psychological Bulletin* 69:161–82. [RMD]

Davidson, E. H. (1987) Understanding embryonic development: A contemporary view. *American Zoologist* 27:581–91. [aDW]

Davison, M. L. & Sharma, A. R. (1988) Parametric statistics and levels of measurement. *Psychological Bulletin* 104:137–44.[aDW]

Debray, Q., Caillard, V. & Stewart. J. (1979) Schizophrenia: A study of genetic models. *Human Heredity* 29:27–36. [aDW]

DeFries, J. C. (1979) Comment. *Theoretical advances in behavior genetics,* ed. J. R. Royce & L. P. Mos. Sijthoff & Noordhoff. [RP]

DeFries, J. C. & Plomin, R. (1978) Behavioral genetics. *Annual Review of Psychology* 29:473–515. [JH]

DeFries, J. C., Wilson, J. R. & McClearn, G. E. (1970) Open-field behavior in mice: Selection response and situational generality. *Behavior Genetics* 1(3):195–211. [H-PL]

Denenberg, V. H. (1977) Interactional effects in early experience research. In:

Genetics, environment and intelligence, ed. A. Oliverio. Elsevier/North Holland Biomedical Press. [aDW]

Detterman, D. K. (1989) The future of intelligence research. *Intelligence* 13(3) 199–204. [DKD]

Dodson, S. (1989) Predator-induced reaction norms. *BioScience* 39:447–52. [AJVN]

Donaldson, T. S. (1968) Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. *Journal of the American Statistical Association* 63:660–76. [DAC]

Donovick, P. J. & Burright, R. G. (1984) Roots to the future: Gene-environment coaction and individual vulnerability to neural insult. In: *Early brain damage, vol. 2, Neurobiology and behavior*, ed. S. Finger & C. R. Almli. Academic Press. [rDW]

Dunn, O. J. & Clark, V. A. (1974) *Applied Statistics: Analysis of variance and regression*. Wiley. [aDW]

Easter, S. S., Jr., Purves, D., Rakic, P. & Spitzer, N. C. (1985) The changing view of neural specificity. *Science* 230:507–11. [aDW]

Eaves, L. J., Last, K., Martin, N. G. & Jinks, J. L. (1977) A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology* 30:1–42. [aDW]

Edwards, A. L. (1979) *Multiple regression and the analysis of variance*. Freeman. [aDW]

(1985) *Experimental design in psychological research*. Harper & Row. [DAC]

Emerson, J. D. & Hoaglin, D. C. (1983) Analysis of two-way tables by medians, In: *Understanding robust and exploratory data analysis*. Wiley. [CG]

Emerson, J. D. & Stoto, M. A. (1983) Transforming data, In: *Understanding robust and exploratory data analysis*. Wiley. [CG]

Erlenmeyer-Kimling, L. (1972) Genotype-environment interactions and the variability of behavior. In: *Genetics, environment, and behavior*, ed. L. Ehrman, G. S. Omenn & E. Caspari. Academic Press. [arDW]

Eysenck, H. J. (1973) *The measurement of intelligence*. Williams and Wilkins. [PHS]

Ewens, W. J. (1979) *Mathematical population genetics*. Springer. [RMD]

Ezekiel, M. & Fox, K. A. (1959) *Methods of correlation and regression analysis*, 3d ed. Wiley. [GMH]

Falconer, D. (1981) *Introduction to quantitative genetics*. Longman Press. [JMC, CJG, rDW]

Fancher, R. E. (1985) *The intelligence men: Makers of the IQ controversy*. Norton. [aDW]

Farber, S. L. (1981) *Identical twins reared apart*. Basic Books. [RMD]

Farmer, A. E., McGuffin, P. & Gottesman, I. I. (1987) Scrutinising the validity of the definition. *Archives of General Psychiatry* 44:634–41. [PM]

Feingold, A. (1988) Cognitive gender differences are disappearing. *American Psychologist* 43:95–103. [rDW]

Finlay, B. M., Wikler, K. C. & Sengelaub, D. R. (1987) Regressive events in brain development and scenarios for vertebrate brain evolution. *Brain, Behavior and Evolution* 30:102–17. [H-PL]

Fisher, R. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52:399–433. [JMC]

Fisher, R. A. (1930) *The genetical theory of natural selection*. Clarendon Press. [JFC]

(1951) Limits to intensive production in animals. *British Agricultural Bulletin* 4:217–18. [JH]

(1958) *The genetical theory of natural selection*, 2d ed. Dover. [CJG]

Fisher, R. A. & Mackenzie, W. A. (1923) Studies in crop variation. II. The manurial responses of different potato varieties. *Journal of Agricultural Science* 13:311–20. [aDW, RMD, OK]

Flechsig, P. (1920) *Anatomie des menschlichen Gehirns und Rückenmarks auf myelogenetischer Grundage. Bd. 1, I Teil. Georg Thiem.* [H-PL]

Flynn, J. R. (1987) Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin* 101: 343–62. [PHS]

Fox, J. (1984) *Linear statistical models and related methods, with applications to social research*. Wiley. [CG]

Freeman, G. H. (1973) Statistical methods for the analysis of genotype-environment interactions. *Heredity* 31:339–54. [aDW, RMD]

French, J. W. (1951) The description of aptitude and achievement tests in terms of rotated factors. *Psychometric monographs*, No. 5. University of Chicago Press. [PHS]

Frigeri, L. G., Wolff, G. L. & Teguh, C. (1988) Differential responses of yellow A^{vy}/A and agouti A/a (BALB/c X VY) F_1 hybrid mice to the same diets: Glucose tolerance, weight gain, and adipocyte cellularity. *International Journal of Obesity* 12:305–20. [rDW]

Fulker, D. W. (1970) Maternal buffering of rodent genotype responses to stress: A complex genotype-environment interaction. *Behavior Genetics* 1:119–124. [GMH]

Fuller, J. L. (1960) Behavior genetics. *Annual Review of Psychology* 11:41–70. [JH]

(1979) Comment. In: *Theoretical advances in behavior genetics*, ed. J. R. Royce & L. P. Mos. Sijthoff & Noordhoff. [NDH]

Fuller, J. L. & Thompson, W. R. (1978) *Foundations of behavior genetics*. Mosby. [aDW]

Futuyma, D. J. (1986) *Evolutionary biology*. Sinauer Associates, Inc. [DAC]

Gaito, J. (1980) Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin* 87:564–67. [aDW]

Gardinier, M. V. & Macklin, W. B. (1988) Myelin proteolipid protein gene expression in jimpy and jimpy^{msd} mice. *Journal of Neurochemistry* 51:360–69. [rDW]

Garner, D. M. & Garfinkel, P. E. (1979). The eating attitude test: An index of the symptoms of anorexia nervosa. *Psychological Medicine* 9:273–80. [PM]

Gerhart, J. C. (1982) The cellular basis of morphogenetic change. Group report. In: *Evolution and development*, ed. J. T. Bonner. Springer-Verlag. [aDW]

Geweke, J. F. & Singleton, K. J. (1980) Interpreting the likelihood ratio test in factor models when sample size is small. *Journal of the American Statistical Association* 75:133–37. [PHS]

Ginsburg, B. E. (1967) Genetic parameters in behavioral research. In: *Behavior-genetic analysis*, ed. J. Hirsch. McGraw-Hill. [rDW]

Ginsburg, B. & Allee, W. C. (1942) Some effects of conditioning on social dominance and subordination in inbred strains of mice. *Physiological Zoology* 15:485–506. [rDW]

Glashow, S. (1988) Tangled in superstring: Some thoughts on the predicament physics is in. *The Sciences* May/June:22–25. [HN]

Gleick, J. (1987) *Chaos*. Viking Press. [rDW]

Goldberger, A. S. (1973) Structural equation models: An overview. In: *Structural equation models in the social sciences*, ed. A. S. Goldberger & O. D. Duncan. Seminar Press. [GMH]

(1978) The nonresolution of IQ inheritance by path analysis. *American Journal of Human Genetics* 39:442–45. [rDW]

Gollin, E. S. (1965) A developmental approach to learning and cognition. In: *Advances in child development and behavior*, vol. 2, ed. L. P. Lipsitt & C. C. Spiker. Academic Press. [DAC]

(1985) Ontogeny, phylogeny, and causality. In: *The comparative development of adaptive skills: Evolutionary implications*, ed. E. S. Gollin. Erlbaum. [DAC]

Goodall, G. & Guastavino, J.-M. (1986) The neurological mutant: Scope and limitations as a tool for the genetic analysis of behaviour. In: *Genetic approaches to behaviour*, ed. J. Medioni & G. Vaysse. I.E.C. [aDW]

Goodlett, C. R., Gilliam, D. M. & West, J. R. (1987) Differential susceptibility of long-sleep (LS) and short-sleep (SS) mice to brain weight reduction following prenatal alcohol exposure. Poster presented at the Behavior Genetics Association annual meeting. Minneapolis, June 25. [rDW]

Gottesman, I. I. & Shields, J. (1982) *Schizophrenia: The epigenetic puzzle*. Cambridge University Press. [PM]

Gould, S. J. (1981) *The mismeasure of man*. Norton. [CG]

Graf, S. A. (1987) The rover/sitter *Drosophila* foraging polymorphism as a function of larval development and food availability. Paper presented at the Behavior Genetics Association annual meeting, Minneapolis, June 25. [rDW]

Grebogi, C., Ott, E. & Yorke, J. A. (1987) Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics. *Science* 238:632–38. [rDW]

Greene, E. (1989) A diet-induced developmental polymorphism in a caterpillar. *Science* 243:643–46. [rDW]

Gregorious, H.-R. & Namkoong, G. (1987) Resolving the dilemmas of interaction, separability, and additivity. *Mathematical Biosciences* 85:51–69. [rDW]

Grosser, M. (1979) *The discovery of Neptune*. Dover. [rDW]

Gupta, A. P. & Lewontin, R. C. (1982) A study of reaction norms in natural populations of *Drosophila pseudo obscure*. *Evolution* 36:934–38. [DAC]

Guastavino, J.-M. (1988) Ethogenèse de la souris mutante *staggerer*, facteurs epigenetiques et recuperations fonctionelles. *Thèse de doctorat d'État des Sciences*, presenteé a L'Université Paris-Nord. [rDW]

Guttman, L. (1986) The irrelevance of factor analysis for the study of group differences. Submitted to *Behavioral and Brain Sciences*. [PHS]

Gutzke, W. H. N. & Crews, D. (1988) Embryonic temperature determines adult sexuality in a reptile. *Nature* 332:832–34. [rDW]

Haavelmo, T. (1943) The statistical implications of a system of simultaneous equations. *Econometrica* 11:1–12. [GMH]

(1944) The probability approach in econometrics. *Econometrica* 12, supplement. [GMH]
Haldane, J. B. S. (1946) The interaction of nature and nurture. *Annals of Eugenics* 13:197–205. [JH]
Harnad, S. (1989) Publication bias: What is its magnitude and nature? Talk delivered at the First International Congress on Peer Review in Biomedical Publication. Chicago, May 10/12, 1989. [PHS]
Harrington, G. M. (1988) Two forms of minority-group test bias as psychometric artifacts with an animal model (*Rattus norvegicus*). *Journal of Comparative Psychology* 102:400–7. [JH]
Hays, W. L. (1988) *Statistics*. 4th ed. Holt, Rinehart & Winston. [aDW]
Heath, A. C., Berg, K., Eaves, L. J., Solaas, M. H., Corey, L. A., Sundet, J., Magnus, P. & Nance, W. E. (1985) Educational policy and the heritability of educational attainment. *Nature* 314:734–36. [aDW]
Hegmann, J. P. & Possidente, B. (1981) Estimating genetic correlations from inbred strains. *Behavior Genetics* 11:103–114. [aDW]
Heikkila, J. J., Browder, L. W., Gedamu, L., Nickells, R. W. & Schultz, G. A. (1986) Heat-shock gene expression in animal embryonic systems. *Canadian Journal of Genetics and Cytology* 28:1093–1105. [aDW]
Henderson, N. D. (1967) Prior treatment effects on open field behaviour of mice: A genetic analysis. *Animal Behaviour* 15:364–76. [RP]
(1968) The confounding effects of genetic variables in early experience research: Can we ignore them? *Developmental Psychobiology* 1:146–52. [NDH]
(1970) Genetic influences on the behavior of mice can be obscured by laboratory rearing. *Journal of Comparative and Physiological Psychology* 73:505–11. [RP]
(1972) Relative effects of early rearing environment on discrimination learning in house mice. *Journal of Comparative and Physiological Psychology* 79:243–53. [RP]
(1979a) Adaptive significance of animal behavior: The role of gene-environment interaction. In: *Theoretical advances in behavior genetics*, ed. J. R. Royce & L. P. Mos. Sijthoff & Noordhoff. [aDW]
(1979b) Reply to comments. In: *Theoretical advances in behavior genetics*, ed. J. R. Royce & L. P. Mos. Sijthoff & Noordhoff. [NDH]
(1982) Human behavior genetics. *Annual Review of Psychology* 33:403–40. [arDW, JH, PM]
(1986) Predicting relationships between psychological constructs and genetic characters: An analysis of changing genetic influences on activity in mice. *Behavior Genetics* 16:201–20. [NDH, RP]
Henry, K. R. (1986) Audiogenic seizures in relation to genetically and experimentally produced cochlear pathology. In: *Perspectives in behavior genetics*, ed. J. L. Fuller & E. C. Simmel. Erlbaum. [aDW]
Heth, C. D., Pierce, W. D., Belke, T. W. & Hensch, S. A. (1989) The effect of logarithmic transformation on estimating the parameters of the generalized matching law. *Journal of the Experimental Analysis of Behavior* 52:65–76. [aDW]
Hewitt, J. K., Eaves, L. J., Neale, M. C. & Meyer, J. M. (1988) Resolving causes of developmental continuity of "tracking." I. Longitudinal twin studies during growth. *Behavior Genetics* 18:133–51. [NDH]
Hill, W. G. (1970) Design of experiments to estimate heritability by regression of offspring on selected parents. *Biometrics* 26:565–71. [rDW]
Hirsch, J. (1981) To "unfrock the charlatans." *Sage Race Relations Abstracts* 6(2):1–65. [JH, PSH, rDW]
(1970) Behavior-genetic analysis and its biosocial consequences. *Seminars in Psychiatry* 2:89–105. [JH]
Ho, M.-W. & Fox, S. W., eds. (1988) *Evolutionary processes and metaphors*. Wiley. [rDW]
Hoaglin, D. C., Mosteller, F. & Tukey, J. W. ed. (1983) *Understanding robust and exploratory data analysis*. Wiley. [CG]
(1985) *Exploring data tables, trends, and shapes*. Wiley. [CG]
Hogben, L. (1939) *Nature and nurture*. W. W. Norton & Co., Inc. (Originally published 1933.) [JH]
(1951) The formal logic of the nature-nurture issue, *Acta Genetica et Statistica Medica* 2:101–40. [JH]
Holland, A. J., Sicotte, N. & Treasure, J. (1988) Anorexia nervosa: Evidence for a genetic basis. *Journal of Psychosomatic Research* 32:561–71. [PM]
Horn, J. M., Loehlin, J. C. & Willerman, L. (1979) Intellectual resemblance among adoptive and biological relatives: The Texas adoption project. *Behavior Genetics* 9:177–207. [RMD]
Hull, C. L. (1945) The place of innate individual and species differences in a natural-science theory of behavior. *Psychological Review* 52:55–60. [aDW]
Huxley, J. (1932) *Problems of Relative Growth*. Dover Press. [JMC]
Hyde, J. S. (1973) Genetic homeostasis and behavior: Analysis, data, and theory. *Behavior Genetics* 3:233–45. [rDW]
(1981) How large are cognitive gender differences? *American Psychologist* 36:892–901. [rDW]
Ingham, P. W. (1988) The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature* 335:25–34. [rDW]
Jacquard, A. (1983) Heritability: One word, three concepts. *Biometrics* 39:465–77. [RMD]
Jencks, C. N., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B. & Michelson, S. (1972) *Inequality*. Harper & Row. [RMD]
Jensen, A. R. (1980) Precis of bias and mental testing, with commentary. *Behavioral and Brain Sciences* 3:325–71. [CG]
Jinks, J. L. & Broadhurst, P. L. (1974) How to analyse the inheritance of behaviour in animals: The biometrical approach. In: *The genetics of behavior*, ed. J. H. F. van Abeelen. North-Holland. [aDW]
Jinks, J. L. & Fulker, D. W. (1970) A comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin* 73:311–49. [aDW, NDH, JH, PHS]
Judd, C. M. & McClelland, G. H. (1989) *Data analysis: A model-comparison approach*. Harcourt Brace Jovanovich. [SEM]
Kamin, L. J. (1974) *The science and politics of IQ*. Erlbaum. [RMD, PHS]
Katz, M. J. (1982) Ontogenetic mechanisms: The middle ground of evolution. In: *Evolution and development*, ed. T. J. Bonner. Springer. [H-PL]
Katz, M. J. & Lasek, R. J. (1978) Evolution of the nervous system: Role of ontogenetic mechanisms in the evolution of matching populations. *Proceedings of the National Academy of Science USA* 75:1349–52. [H-PL]
Keightley, P. D. (1989) Models of quantitative variation of flux in metabolic pathways. *Genetics* 121:869–976. [JFC]
Kempthorne, O. (1952) *The design and analysis of experiments*. Wiley. [aDW, OK]
(1957) An introduction to genetic statistics. Wiley. Reprinted 1969. Iowa State University Press. [OK]
(1978) Logical, epistemological and statistical aspects of nature-nurture data interpretation. *Biometrics* 34:1–23. [OK]
Kendell, R. E. (1982) The choice of diagnostic criteria for biological research. *Archives of General Psychiatry* 39:1334–39. [PM]
Kendler, K. S. & Eaves, L. J. (1986) Models for the joint effect of genotype and environment on liability to psychiatric illness. *American Journal of Psychiatry* 143:279–89. [aDW]
Kennedy, J. L., Giuffra, L. A., Moises, H. W., Cavalli-Sforza, L. L., Pakstis, A. J., Kidd, J. R., Castiglione, C. M., Sjogren, B., Wetterberg, L. & Kidd, K. K. (1988) Evidence against linkage of schizophrenia to markers on chromosome 5 in a northern Swedish pedigree. *Nature* 336:167–70. [H-PL]
Kenny, D. A. (1987) *Statistics for the social and behavioral sciences*. Little, Brown. [SEM]
Kerner, A.-L. & Carson, J. H. (1986) Shiverer*jimpy double mutant mice. I. Biochemical evidence for reciprocal intergenic suppression. *Brain Research* 374:45–53. [rDW]
Kevles, D. J. (1985) *In the name of eugenics*. Knopf. [aDW]
Killian, A. M. (1989) Playing dice with the solar system. *Sky and Telescope* 78 (Aug.):136–40. [rDW]
Kinsley, C. & Svare, B. (1987) Genotype modulates prenatal stress effects on aggression in male and female mice. *Behavioral and Neural Biology* 47:138–150. [aDW]
Klein, D. C. & Yuwiler, A. (1973) β-adrenergic regulation of indole metabolism in the pineal gland. In: *Frontiers in catecholamine research*, ed. E. Usdin & S. H. Snyder. Pergamon Press. [aDW]
Koele, P. (1982) Calculating power in analysis of variance. *Psychological Bulletin* 92:513–16. [rDW]
Kraemer, H. C. & Thiemann, S. (1987) *How many subjects? Statistical power analysis in research*. Sage. [aDW]
Kurnit, D. M., Layton, W. M. & Matthysse, S. (1987) Genetics, chance, and morphogenesis. *American Journal of Human Genetics* 41:979–95. [rDW]
Kvålseth, T. O. (1985) Cautionary note about R^2. *American Statistician* 39:279–85. [aDW]
Lachenbruch, P. A. (1988) A note on sample size computations for testing interactions. *Statistics in Medicine* 7:467–69. [aDW]
Lamoreux, M. L. & Pendergast, P. (1987) Genetic controls over melanocyte differentiation: Interaction of agouti-locus and albino-locus genetic defects. *The Journal of Experimental Zoology* 243:71–79. [rDW]
Lande, R. (1976) Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30:314–34. [JMC]
(1979) Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution* 33:402–17. [JMC]
Lassalle, J. M. (1986) Les interactions entre génotype et environment. *Psychologie Française* 31:205–211. [aDW]

Lathrope, G. M., Lalouel, J. M. & Jacquard, A. (1984) Path analysis of family resemblance and gene-environment interaction. *Biometrics* 40:611–625. [aDW]

Layzer, D. (1974) Heritability analyses of IQ scores: Science or numerology? *Science* 183:1259–66. [RMD]

Leahy, A. M. (1935) Nature-nurture and intelligence. *Genetic Psychology Monographs* 17(4):236–307. [RMD]

Lemke, G. (1988) Unwrapping the genes for myelin. *Neuron* 1:535–43. [rDW]

Le Roy, H. L. (1960a) *Statistische methoden der populationsgenetik.* Birkhäuser. [RMD]

(1960b) The interpretation of calculated heritability coefficients with regard to gene and environmental effects as well as to genotype-environment interactions. In: *Biometrical genetics*, ed. O. Kempthorne. Pergamon. [RMD]

Levin, J. R. (1975) Determining sample size for planned and *post hoc* analysis of variance comparisons. *Journal of Educational Measurement* 12:99–108. [SEM]

Levins, R. & Lewontin, R. (1985) *The dialectical biologist.* Harvard University Press. [aDW]

Lewontin, R. (1974) The analysis of variance and the analysis of causes. *American Journal of Human Genetics* 26:400–411. [aDW, DAC]

(1982) Organism and environment. In: *Learning, development, and culture*, ed. H. C. Plotkin. Wiley. [rDW]

Lichtenstein, G. (1977) Fund backs controversial study of 'racial betterment'. *The New York Times*, December 11, p. 1. [rDW]

Lindzey, G., Loehlin, J., Manosevitz, M. & Thiessen, D. (1971) Behavioral genetics. *Annual Review of Psychology* 22:39–94. [JH]

Lipp, H.-P. (1979) Brain complexity enhances speed of behavioral evolution. *Behavioral and Brain Sciences* 2:42. [H-PL]

(1989) Non-mental aspects of encephalisation: The forebrain as a playground of mammalian evolution. *Human Evolution.* [H-PL]

Lipp, H.-P. & Schwegler, H. (1982) Hippocampal mossy fibers and avoidance learning. In: *Genetics of the Brain*, ed. I. Lieblich. Elsevier. [H-PL]

Lipp, H.-P., Schwegler, H., Crusio, W. E., Wolfer, D. P., Heimrich, B., Driscoll, P. & Leisinger-Trigona, M.-C. (1989) Using genetically defined rodent strains for the identification of hippocampal traits relevant for two-way avoidance learning: A noninvasive approach. *Experientia* 45:845–59. [H-PL]

Lipp, H.-P, Schwegler, H., Heimrich, B. & Driscoll, P. (1988) Infrapyramidal mossy fibers and two-way avoidance learning: Developmental modification of hippocampal circuitry and adult behavior of rats and mice. *The Journal of Neuroscience* 8:1905–21. [H-PL]

Loehlin, J. C., Willerman, L. & Horn, J. M. (1988) Human behavior genetics. *Annual Review of Psychology* 39:101–33. [JH-rDW]

Loevinger, J. (1943) On the proportional contributions of differences in nature and in nurture in differences in intelligence. *Psychological Bulletin* 40(10):725–56. [JH]

Lubin, A. (1961) The interpretation of significant interaction. *Educational and Psychological Measurement* 21:807–817. [aDW]

Lyon, M., Barr, C. E., Cannon, T. D., Mednick, S. A. & Shore, D. (1989) Fetal neural development and schizophrenia. *Schizophrenia Bulletin* 15:149–60. [H-PL]

Mackintosh, N. J. (1974) *The psychology of animal learning.* Academic Press. [aDW]

Maddox, J. (1984) Genetics and hereditable IQ. *Nature* 309:579. [JH]

Magnusson, D. & Allen, V. L. (1983) An interactional perspective for human development. In: *Human development: An international perspective*, ed. D. Magnusson & V. L. Allen. Academic Press. [aDW]

Mandel, J. (1961) Non-additivity in two-way analysis of variance. *American Statistical Association Journal* 56:878–88. [aDW]

Marler, M. R. (1980) Likelihood ratio tests of hypotheses: Comments on Pitz's article and some alternative procedures. *Psychological Bulletin* 87:568–74. [aDW]

Martin, N. G. Boomsma, D. I., Neale, M. C. (1989) Foreword. Special issue: Twin methodology using LISREL. *Behavior Genetics* 19:5–7. [JH]

Martin, N. G., Eaves, L. J., Kearsey, M. J. & Davies, P. (1978) The power of the classical twin study. *Heredity* 40:97–116. [NDH]

Masur, J. & Benedito, M. A. C. (1974) Genetic selection of winner and loser rats in a competitive situation. *Nature* 249:284. [H-PL]

Mather, K. & Jinks, J. L. (1982) *Biometrical genetics: The study of continuous variation.* Chapman & Hall. [aDW]

Maxwell, S. E. & Delaney, H. D. (1989) *Designing experiments and analyzing data: A model comparison's perspective.* Wadsworth. [SEM]

May, R. W. (1960) Genetics and subversion. *Nation*, May 14, 420–2. [rDW]

McClearn, G. E. & Meredith, W. (1966) Behavioral genetics. *Annual Review of Psychology* 17:515–20. [JH]

McGue, M., Gottesman, I. I. & Rao, D. C. (1983) The transmission of schizophrenia under a multifactorial threshold model. *American Journal of Human Genetics* 35:1161–78. [H-PL]

(1985) Resolving genetic models for the transmission of schizophrenia. *Genetic Epidemiology* 2:99–110. [PM]

McGuffin, P., Farmer, A. E., Gottesman, I. I., Murray, R. M. & Reveley, A. M. (1984) Twin concordance for operationally defined schizophrenia: Confirmation of familiality and heritability. *Archives of General Psychiatry* 4:1541–45. [PM]

McGuffin, P., Katz, R. & Aldrich, J. (1986) Past and present state examination: The assessment of "lifetime ever" psychopathology. *Psychological Medicine* 16:461–65. [PM]

McGuffin, P., Katz, R. & Bebbington, P. (1988) The Camberwell collaborative depression study III. Depression and adversity in the relatives of depressed probands. *British Journal of Psychiatry* 152:775–82. [PM]

McGuire, T. R. & Hirsch, J. (1977) General intelligence (g) and heritability (H^2, h^2). In: *The structuring of experience*, ed. I. C. Uzgiris & F. Weizmann. Plenum. [aDW, JH]

Meehl, Paul E. (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46:806–34. [JH]

Moisset, B. (1977) Factors contributing to the modulation of norepinephrine uptake by synaptosomes from mouse brain cortex. *Brain Research* 121:113–20. [rDW]

Moran, P. A. P. (1968) *An introduction to probability theory.* Clarendon Press. [JFC]

Nanney, D. L. (1986) Eugenics and human heredity. *The Journal of Heredity* 77:481–82. [rDW]

Nave, K.-A., Bloom, F. E. & Milner, R. J. (1987) A single nucleotide difference in the gene for myelin proteolipid protein defines the *jimpy* mutation in mouse. *Journal of Neurochemistry* 49:1873–77. [rDW]

Newman, H. H., Freeman, F. N. & Holzinger, K. J. (1937) *Twins: A study of heredity and environment.* University of Chicago Press. [RMD, PHS]

Neyman, J. (1935) Comments on Mr. Yates' paper. *Journal of the Royal Statistical Society, Supplement* 2:235–41. [aDW, OK]

Nordskog, A. W. (1977) Success and failure of quantitative genetic theory in poultry. In: *Proceedings of the International Conference on Quantitative Genetics*, ed. E. Pollak, O. Kempthorne & T. B. Bailey, Jr. Iowa State University Press. [rDW]

Norton, D. W. (1952) An empirical investigation of some effects of non-normality and heterogeneity of the F-distribution. Doctoral dissertation, State University of Iowa. [DAC]

Nosten, M. (1989) Early development in mice VI. Additive and interactive effects of offspring genotype and maternal environments. *Physiology and Behavior* 45:955–61. [MC]

Nosten, M. & Roubertoux, P. L. (1988) Uterine and cytoplasmic effects on pup eyelid opening in two inbred strains of mice. *Physiology and Behavior* 43:167–71. [MC]

Nyborg, H. (1977) *The rod-and-frame test and the field dependence dimension: Some methodological, conceptual, and developmental considerations.* Dansk Psykologisk Forlag. [HN]

(1983) Spatial ability in men and women: Review and new theory. *Advances in Behavior Research and Therapy* (Monograph Series) 5(whole no. 2):89–140. [HN]

(1984) Performance and intelligence in hormonally-different groups. In: *Sex differences in the brain: The relation between structure and function*, ed. G. J. De Vries, J. P. C. de Bruin, H. B. M. Uylings & M. A. Corner. Progress in Brain Research, vol. 61. Elsevier Biomedical Press. [HN]

(1987) Individual differences or different individuals? That is the question. *Behavioral and Brain Sciences* 10:34–35. [HN]

(1988) Mathematics, sex hormones, and brain function. *Behavioral and Brain Sciences* 11:206–7. [HN]

(1989) Sex hormones, brain development, and spatio-perceptual strategies in women with Turner's syndrome and in school girls. In: *Sex chromosome abnormalities and behavior: Psychological studies*, ed. B. Bender & D. Berch. Westview Press. [HN]

(submitted a) Sex, body, mind, and society: A physicological approach. [HN]

(submitted b) Nonlinear harmonization of body, brain, and intellectual development: A model, a theory, and a research program. [HN]

Nyborg, H. & Boeggild, C. (1989) Origin of individual and sex differences in body and ability. Paper presented at the fourth meeting of the International Society for the Study of Individual Differences. Federal Republic of Germany, June 22–25. [HN]

Oliverio, A., Eleftheriou, B. E. & Bailey, D. W. (1973) Exploratory activity:

Genetic analysis of its modification by scopolamine and amphetamine. *Physiology and Behavior* 10:893–99. [rDW]
Oyama, S. (1985) *The ontogeny of information*. Cambridge University Press. [aDW, DB]
(1988) Reply to Robert Plomin's review of *The ontogeny of information*. *Developmental Psychobiology* 21:97–100. [aDW]
Palmer, A. R. & Strobeck, C. (1986) Fluctuating asymmetry: Measurement, analysis, patterns. *Annual Review of Ecology and systematics* 17:391–421. [rDW]
Parsons, P. A. (1988) Behavior, stress and variability. *Behavior Genetics* 18:293–308. [NDH]
Pearson, K. A. (1930) *Tables for statisticians and biometricians*. Cambridge University Press. [GMH]
Peeler, D. F. (1986) Circadian variations in activity of neuroanatomically distinct recombinant inbred mice as a function of genetic and measurement variables. *Society for Neuroscience Abstracts* 12:1069. [rDW]
Peeler, D. F. & Nowakowski, R. S. (1987) Active avoidance performance in genetically defined mice. *Behavioral and Neural Biology* 48:83–89. [rDW]
Perkins, J. M. & Jinks, J. L. (1973) The assessment and specificity of environmental and genotype-environmental components of variability. *Heredity* 30:111–126. [aDW]
Peters, D. P. & Ceci, S. J. (1982) Peer review practices of psychological journals: The fate of articles, submitted again. *Behavioral and Brain Sciences* 5:187–255. [PHS]
Phillips, K., Fulker, D. W. & Rose, R. J. (1987) Path analysis of seven fear factors in adult twin and sibling pairs and their parents. *Genetic Epidemiology* 4:345–55. [aDW]
Pike, W. L. (1978) *Short-term instruction, test-wiseness, and the Scholastic Aptitude Test*. College Entrance Examination Board. [PHS]
Platt, S. A. & Sanislow, C. A., III (1988) The norm-of-reaction: Definition and misinterpretation of animal research. *Journal of Comparative Psychology* 102:254–61. [aDW]
Plomin, R. (1983) Developmental behavior genetics. *Child Development* 54:253–59. [RMD]
(1986) *Development, genetics and psychology*. Erlbaum. [aDW, DKD, RL]
(1988) Reply to Susan Oyama's review of *Development, genetics and Psychology*. *Development Psychobiology* 21:107–12. [aDW, RL]
Plomin, R. & Daniels, D. (1987) Why are children in the same family so different from one another? *Behavioral & Brain Sciences* 10:1–60. [WEC, rDW]
Plomin, R. & DeFries, J. C. (1983) The Colorado adoption project. *Child Development* 54:276–89. [aDW]
Plomin, R., DeFries, J. C. & Fulker, D. W. (1988) *Nature and nurture during infancy and early childhood*. Cambridge University Press. [RL]
Plomin, R., DeFries, J. C. & Loehlin, J. C. (1977) Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin* 84:309–22. [aDW, RMD, JH, RL]
Plomin, R., DeFries, J. C. & McClearn, G. E. (1980) *Behavioral genetics, a primer*. Freeman. [aDW, WEC, RL]
Plomin, R., Loehlin, J. C. & DeFries, J. C. (1985) Genetic and environmental components of "environmental" influences. *Developmental Psychology* 21:391–402. [aDW]
Poston, T. & Stewart, I. (1978) *Catastrophe theory and its applications*. Pitman Publishing, Ltd. [FLB]
Powers, L. (1950) Determining scales and the use of transformations in studies of weight per locule of tomato fruit. *Biometrics* 6:145–63. [GMH]
(1955) Components of variance method and partitioning method of genetic analysis applied to weight per fruit of tomato hybrid and parental population. *Bulletin 1131 of the U.S. Department of Agriculture*. [GMH]
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988) *Numerical recipes in C*. Cambridge University Press. [rDW]
Pritchard, D. J. (1986) *Foundations of developmental genetics*. Taylor & Francis. [aDW, HN]
Provine, W. (1971) *The origins of theoretical population genetics*. University of Chicago Press. [JMC]
Reid, J. B. & Murfet, I C. (1977) Flowering in *Pisum:* The effect of light quality on the genotype *If e Sn Hr*. *Journal of Experimental Botany* 28:1357–64. [rDW]
Reik, W., Collick, A., Norris, M. L., Barton, S. C. & Surani, A. (1987) Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature* 328:248–51. [rDW]
Ricker, J. P. & Hirsch, J. (1988) Reversal of genetic homeostasis in laboratory populations of *Drosophila melanogaster* under long-term selection for geotaxis and estimates of gene correlates: Evolution of behavior-genetic systems. *Journal of Comparative Psychology* 102:203–14. [rDW]
Riska, B. (1986) Some models for development, growth, and morphometric correlation. *Evolution* 40:1303–11. [JMC]
Roberts, D. W. A. & Larson, R. I. (1985) Vernalization and photoperiodic responses of selected chromosome substitution lines derived from Rescue, Cadet, and Cypress wheats. *Canadian Journal of Genetics and Cytology* 27:586–91. [rDW]
Rosenthal, R. & Rosnow, R. L. (1985) *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge University Press. [SEM]
Rotton, J. & Schönemann, P. H. (1978) Power tables for the analysis of variance. *Educational and Psychological Measurement* 38:213–29. [rDW]
Rouanet, H., Lepline, D. & Holender, D. (1978) Model acceptability and the use of Bayes-fiducial methods for validating models. In *Attention and Performance VII*, ed. J. Requin. Erlbaum. [MC]
Rouanet, H., Lecoutre, B. (1983) Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology* 36:252–68. [MC]
Roubertoux, P. L. (1981) Valeur explicative du concept d'interaction en analyse génétique. *Les Niveaux d'Explication en Psychologie*. Colloque CNRS. [MC]
Roubertoux, P. L., Nosten-Bertrand, M. & Carlier, M. (in press) Additive and interactive effects between genotype and environment: Concepts and facts. *Advances in the Study of Behavior*. [MC]
Roux, C. Z. (1984) Treatment x unit interactions in the completely randomized and randomized block designs, In: *Experimental design, statistical models, and genetic statistics*, ed. K. Hinkelmann. Marcel Dekker. [aDW]
Rowe, D. C. (1987) Resolving the person-situation debate. *American Psychologist* 42:218–27. [aDW]
Rutherford, J., Katz, R. & McGuffin, P. (in preparation) Genes, environment, and attitudes to eating. [PM]
Salsburg, D. S. (1985) The religion of statistics as practiced in medical journals. *The American Statistician* 39:220–23. [rDW]
Sapienza, C., Peterson, A. C., Rossant, J. & Balling, R. (1987) Degree of methylation of transgenes is dependent on gamete of origin. *Nature* 328:251–54. [rDW]
Satinder, K. P. & Sterling, J. W. (1983) Differential effects of pre- and/or post-natal *d*-amphetamine on avoidance response in genetically selected lines of rats. *Neurobehavioral Toxicology and Teratology* 5:315–20. [rDW]
Scarr, S. & Weinberg, R. A. (1976) IQ test performance of black children adopted by white families. *American Psychologist* 31:726–39. [RMD]
Scharloo, W. (1989) Developmental and physiological aspects of reaction norm. *BioScience* 39:465–71. [AJVN]
Scheffé, H. (1959) *The analysis of variance*. Wiley. [aDW, PHS]
Schiff, M., Duyme, M., Dumaret, A., Stewart, J., Tomkiewicz, S. & Feingold, J. (1978) Intellectual status of working-class children adopted early into upper middle class families. *Science* 200:1503–4. [RMD]
Schlichting, C. D. (1989) Phenotypic integration and environmental change. *BioScience* 39:460–64. [AJVN]
Schmalhausen, I. I. (1949) *Factors of evolution*. Blakiston. Reprinted 1986 by University of Chicago Press. [aDW]
Schneider, K. (1959) *Clinical psychopathology*. Translated by M. W. Hamilton. Grune and Stratton. [PM]
Schönemann, P. H. (1981) Power as a function of communality in factor analysis. *Bulletin of the Psychonomic Society* 17:57–60. [PHS]
(1987) Letter to J. Hirsch, November 30, 1987. In J. Hirsch papers, University of Illinois Archives, No. 15/19/22. [JH]
(1989a) New questions about old heritability estimates. *Bulletin of the Psychonomic Society* 27:175–78. [PHS]
(1989b) Some new results on the Spearman hypothesis artifact. *Bulletin of the Psychonomic Society*. [PHS]
Schulsinger, F., Parnas, J., Mednick, S., Teasdale, T. W. & Schulsinger, H. (1987) Heredity-environment interaction and schizophrenia. *Journal of Psychiatric Research* 21:431–36. [rDW]
Scott, J. P. (1942) Genetic differences in the social behavior of inbred strains of mice. *The Journal of Heredity* 33:11–15. [rDW]
Sedlmeier, P. & Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105:309–16. [SEM]
Severo, N. C. & Zelen, M. (1960) Normal approximation to the chi-square and non-central F probability functions. *Biometrika* 47:411–16. [aDW]
Shields, J. (1960) *Monozygotic twins brought up apart and brought up together*. Oxford University Press. [PHS]
Shockley, W. (1987) Jensen's data on Spearman's hypothesis: No artifact. *Behavioral and Brain Sciences* 10:512. [PHS]

Simmel, E. C. & Eleftheriou, B. E. (1977) Multivariate and behavior genetic analysis of avoidance of complex visual stimuli and activity in recombinant inbred strains of mice. *Behavior Genetics* 7:239–50. [rDW]

Skodak, M. & Skeels, H. M. (1949) A final followup study of one hundred adopted children. *Journal of Genetic Psychology* 75:85–125. [RMD]

Slack, W. V. & Porter, D. (1980) The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review* 50:154–75. [PHS]

Slatkin, M. (1987) Quantitative genetics of heterochrony. *Evolution* 41:799–811. [JMC]

Smotherman, W. P. & Robinson, S. R. (1988) Behavior of rat fetuses following chemical or tactile stimulation. *Behavioral Neuroscience* 102:24–34. [rDW]

Snedecor, G. W. & Cochran, W. G. (1980) *Statistical methods.* Iowa State University Press. [CG]

(1967) Statistical methods. Iowa State University Press. [PHS]

Soper, H. V., Cicchetti, D. V., Satz, P., Light, R. & Orsini, D. L. (1988) Null hypothesis disrespect in neuropsychology: Dangers of alpha and beta errors. *Journal of Clinical and Experimental Neuropsychology* 10:255–70. [DVC]

Stearns, S. C. (1989) The evolutionary significance of phenotypic plasticity. *BioScience* 39:436–446. [AJV N]

Stent, G. S. (1981) Strength and weakness of the genetic approach to the development of the nervous system. *Annual Review of Neuroscience* 4:163–94. [aDW]

Stein, G. J. (1988) Biological science and the roots of Nazism. *American Scientist* 76:50–58. [rDW]

Tang, P. C. (1938) The power function of the analysis of variance tests with tables and illustrations of their use. *Statistical Research Memoirs* 2:126–57. [aDW]

Taylor, E. E. & Condra, C. (1978) Genetic and environmental interaction in *Drosphila pseudoobscura*. *Journal of Heredity* 69:63–64. [RP]

Taylor, H. F. (1980) *The IQ game.* Rutgers University Press. [rDW]

Tienari, P., Sorri, A., Lahti, I., Naarala, M., Wahlberg, K.-E., Moring, J., Pohjola, J. & Wynne, L. C. (1987) Genetic and psychosocial factors in schizophrenia: The Finnish adoptive family study. *Schizophrenia Bulletin* 13:477–84. [rDW]

Tiku, M. L. (1966) A note on approximating to the noncentral F distribution. *Biometrika* 53:606–10. [aDW]

(1971) Power function of the F-test under non-normal situations. *Journal of the American Statistical Association* 66:913–16. [DAC]

Townsend, J. T. & Ashby, F. G. (1984) Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin* 96:394–401. [aDW]

Traxler, R. H. (1976) A snag in the history of factorial experiments. In: *On the history of statistics and probability,* ed. D. B. Owen. Marcel Dekker. [aDW, DVC]

Tukey, J. W. (1957) On the comparative anatomy of transformations. *Annals of Mathematical Statistics* 28:602–32. [aDW]

Van Abeelen, J. H. F., Van der Kroon, P. H. & Bekkers, M. F. (1973) Mice selected for rearing behavior: Some physiological variables. *Behavior Genetics* 3(1):85–90. [H-PL]

Van Noordwijk, A. J. (1989) Reaction norms in genetical ecology. *BioScience* 39:453–59. [AJVN]

Van Oortmerssen, G. A. & Bakker, T. C. (1981) Artificial selection for short and long attack latencies in wild Mus musculus domesticus. *Behavior Genetics* 11(2):115–26. [H-PL]

Vardimon, L., Fox, L. L., Degenstein, L. & Moscona, A. A. (1988) Cell contacts are required for induction by cortisol of glutamine synthetase gene transcription in the retina. *Proceedings of the National Academy of Science U.S.A.* 85:5981–85. [rDW]

Vetta, A. (1981) Appendix. In: J. Hirsch, To "unfrock the charlatans." *Sage Race Relations Abstracts* 6:35–39. [aDW, PHS]

Via, S. (1984a) The quantitative genetics of polyphagy in an insect herbivore. I. Genotype-environment interaction in larval performance on different host plant species. *Evolution* 38:881–95. [CJG]

(1984b) The quantitative genetics of polyphagy in an insect herbivore. II. Genetic correlations in larval performance within and among host plants. *Evolution* 38:896–905. [CJG]

Via, S. & Lande, R. (1985) Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution* 39:505–22. [aDW, JMC, CJG]

Wachs, T. D. (1983) The use and abuse of environment in behavior-genetic research. *Child Development* 54:396–407. [rDW]

Wahlsten, D. (1978) Behavioral genetics and animal learning. In: *Psychopharmacology of aversively motivated behavior,* ed. H. Anisman & G. Bignami. Plenum. [rDW]

(1979) A critique of the concepts of heritability and heredity in behavioral genetics. In: *Theoretical advances in behavior genetics,* ed. J. R. Royce & L. P. Mos. Sijthoff & Noordhoff. [aDW, NDH]

(1983) Maternal effects on mouse brain weight. *Developmental Brain Research* 9:216–21. [rDW]

(1987a) Insensitivity of analysis of variance to heredity-environment interaction. Paper presented at the Behavior Genetics Association annual meeting, Minneapolis, June 27. [rDW]

(1987b) Three sources of individual differences. Paper presented at the Canadian Psychological Association annual meeting, Vancouver, June. [rDW]

(1989a) Genetic and developmental defects of the mouse corpus callosum. *Experientia.*45:828–38. [rDW]

(1989b) Sample size and power to detect a linear contrast, with application to interaction in a 2 × 2 design. Unpublished manuscript. [rDW]

Wainwright, P. (1980) Relative effects of maternal and pup heredity on post-natal mouse development. *Developmental Psychobiology* 13:493–98. [rDW]

Ward, R. (1985) Genetic polymorphisms and additive genetic models. *Behavior Genetics* 15:537–48. [rDW]

Wilson, L. M. & Sinha, H. L. (1985) Thermal preference behavior of genetically obese (*ob/ob*) and genetically lean (+/?) mice. *Physiology and Behavior* 35:545–58. [rDW]

Wimer, R. E. & Wimer, C. C. (1982) A biometrical-genetic analysis of granule cell number in the area dentata of house mice. *Developmental Brain Research* 2:129–40. [rDW]

Winer, B. J. (1971) *Statistical principles in experimental design,* 2nd ed. McGraw-Hill. [aDW]

Winslow, G. M., Carroll, S. B. & Scott, M. P. (1988) Maternal-effect genes that alter the fate map of the *Drosophila* blastoderm embryo. *Developmental Biology* 129:72–83. [rDW]

Wittgenstein, L. (1953) *Philosophical investigations.* Macmillan. [DB]

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research* 20:557–85. [aDW]

(1934) The method of path coefficients. *Annals of Mathematical Statistics* 5:161–215. [GMH]

(1978) The relation of livestock breeding to theories of evolution. *Journal of Animal Science* 46:1192–1200. [rDW]

Yates, F. (1935) Complex experiments. *Journal of the Royal Statistical Society, Supplement* 2:181–223. [aDW]

Yeakley, J. M., Janavs, J. L. & Reiness, C. G. (1987) Muscle activity pattern regulates postnatal development of acetylcholinesterase molecular forms in normal mice and mice with motor endplate disease. *The Journal of Neuroscience* 7:4084–94. [rDW]

Zacharko, R. M., Lalonde, G. T., Kasian, M. & Anisman, H. (1987) Strain-specific effects of inescapable shock on intercranial self-stimulation from the nucleus accumbens. *Brain Research* 426:164–68. [rDW]

Medical Research Modernization Committee

The MRMC announces the publication of the first volume of *Perspectives on Animal Research,* an annual series of essays and commentaries.

This issue includes articles on the following topics:

Kathryn Hahner -- Learned helplessness

Christopher D. Smith -- Head injury research

Brandon P. Reines -- Physiological psychology

Nedim C. Buyukmihci -- Visual deprivation, ocular histoplasmosis, laser retinotoxicity

Kenneth P. Stoller -- Toxicity testing

Stephen R. Kaufman -- Ocular toxicity, dermal toxicity

Irwin D. Bross -- Mathematical models versus animal models

Gary I. Block -- Animal use in veterinary school

Orders to: MRMC, Box 6036, New York, NY 10163-6018. $15.00 hardcover, $8.00 paperback (134 pp.)

The Medical Research Modernization Committee is composed of health care professionals who lend their experience and expertise to identify and promote modern methods of biomedical research, many of which have exceeded or outdated traditional animal "models" in accuracy and relevance.

https://doi.org/10.1017/S0140525X00078055